Department of Computer Science

Faculty of Computing

**DATA MINING**

**LAB ASSIGNMENT 2 UNSUPERVISED LEARNING MODEL (CLUSTERING)**

| | | |
|---|---|---|
| **Programme** | : | Bachelor of Computer Science (*Data Engineering*) |
| **Subject Code** | : | SECP2753 |
| **Session-Sem** | : | 2023/2024-2 |

**Prepared by** :

- ☐ LOO JIA CHANG (A22EC0074)
- ☐ GOH JING YANG (A22EC0052)

**Section** : 01

**Topic** : Predictive Analytics for Cardiometabolic Risk Factors: A Multifactorial Approach to Forecasting Diabetes, Hypertension, and Stroke Incidences

**Lecturer** : DR. ROZILAWATI BINTI DOLLAH

**Date** : 07/06/2024

Contents

1. Introduction

In the dynamic field of data science, clustering forms an essential part of unsupervised learning, used extensively across diverse applications such as customer segmentation, anomaly detection, and organizing large sets of unlabeled data. This lab assignment focuses on the application of clustering techniques to uncover inherent groupings within data. We will explore different clustering algorithms to understand their utility and effectiveness in revealing patterns and structures without prior labeling of the outcomes.

Objectives

The primary objectives of this lab are to:

- Implement and explore various clustering algorithms to determine which technique best identifies and segregates similar data points into distinct groups within a given dataset.
- Assess the performance of these clustering models using appropriate evaluation metrics and visualizations to understand the quality and practicality of the derived clusters.
- Enhance practical skills in Python programming for machine learning and gain a deeper understanding of the algorithmic underpinnings and challenges associated with unsupervised learning.

Tools and Technologies

In this assignment, the following tools and technologies are utilized:

- Python Version 3.10.14: Chosen for its strong ecosystem of libraries and frameworks that support data manipulation and machine learning.
- Scikit-Learn Version 1.0.2: Provides robust tools for data mining and data analysis, including several pre-implemented clustering algorithms.
- Matplotlib and Seaborn: Used for creating visualizations to analyze the effectiveness of the clustering results.
- Pytorch: Chosen for its strong ecosystem of libraries and frameworks that support data manipulation and machine learning. It is use to driven the K means algorithm in this project.

## Methodology

The methodology for this lab includes several key phases:

1. Data Preprocessing:
   - Standardization: To ensure that each feature contributes equally to the result, we standardize the data, removing mean and scaling to unit variance.
   - Handling Missing Values: Impute missing values, if necessary, to maintain the integrity of the dataset.
2. Clustering Algorithms Implementation:
   - K-Means Clustering: Implement to explore centroid-based clustering.
   - Hierarchical Clustering: Use to investigate agglomerative clustering techniques.
   - DBSCAN: Apply to understand density-based clustering capabilities.
3. Evaluation:
   - Silhouette Score: Measure the quality of clusters formed by different algorithms.
   - Elbow Method: For K-means, determine the optimal number of clusters.
   - Dendrograms: For hierarchical clustering, visualize how clusters are formed at different scales.
4. Visualization:
   - Use scatter plots and color-coded clusters to visualize and interpret the clustering results, providing intuitive insights into the data grouping.
5. Discussion and Analysis:
   - Discuss the results obtained from different clustering techniques and their practical implications.
   - Analyze the challenges encountered during implementation and potential improvements.

Project Link: [23242_DM_Lab2_Clustering/LAB2Clustering at main · jcl03/23242_DM_Lab2_Clustering (github.com)]

2. Dataset

Dataset Link: DATA MINING DATASET.xlsx

3. Characteristic of Data

| Columns | Description | Types of Attributes |
|---|---|---|
| Age | 13-level age category (_AGEG5YR see codebook) 1 = 18-24 9 = 60-64 13 = 80 or older | Ordinal |
| Sex | Patient's gender (1: male; 0: female) | Binary |
| HighChol | 0 = no high cholesterol 1 = high cholesterol | Binary |
| CholCheck | 0 = no cholesterol check in 5 years 1 = yes cholesterol check in 5 years | Binary |
| BMI | Body Mass Index | Ordinal |
| Smoker | Have you smoked at least 100 cigarettes in your entire life? [Note: 5 packs = 100 cigarettes] 0 = no 1 = yes | Binary |
| HearthDiseaseorAttack | coronary heart disease (CHD) or myocardial infarction (MI) 0 = no 1 = yes | Binary |
| PhysActivity | physical activity in past 30 days (about 4 and a half weeks) - not including job 0 = no 1 = yes | Binary |
| Fruits | Consume Fruit 1 or more times per day 0 = no 1 = yes | Binary |

| | | |
|---|---|---|
| Veggies | Consume Vegetables 1 or more times per day 0 = no 1 = yes | Binary |
| HvyAlcoholComsump | (Adult men >=14 drinks per week and adult women>=7 drinks per week) 0 = no 1 = yes | Binary |
| GenHlth | Would you say that in general your health is: scale 1-5 1 = excellent 2 = very good 3 = good 4 = fair 5 = poor | Ordinal |
| MentHlth | Days of poor mental health scale 1-30 days (about 4 and a half weeks) | Ordinal |
| PhysHlth | Physical illness or injury days in past 30 days (about 4 and a half weeks) scale 1-30 | Ordinal |
| DiffWalk | Do you have serious difficulty walking or climbing stairs? 0 = no 1 = yes | Binary |
| Stroke | you ever had a stroke. 0 = no, 1 = yes | Binary |
| HighBP | 0 = no high, BP 1 = high BP | Binary |
| Diabetes | 0 = no diabetes, 1 = diabetes | Binary |

Data Preprocessing

1. Missing value handling

```python
#load data
data = pd.read_csv('data/diabetes_data.csv')

#check for missing values
print(data.isnull().sum())
```

```
Age                     0
Sex                     0
HighChol                0
CholCheck               0
BMI                     0
Smoker                  0
HeartDiseaseorAttack    0
PhysActivity            0
Fruits                  0
Veggies                 0
HvyAlcoholConsump       0
GenHlth                 0
MentHlth                0
PhysHlth                0
DiffWalk                0
Stroke                  0
HighBP                  0
Diabetes                0
dtype: int64
```

The dataset does not have any missing value to be handle

2. Remove duplicates

```python
#check for duplicates
print(data.duplicated().sum())
#drop duplicates
data.drop_duplicates(inplace=True)
print(data.info())

#show min and max values for each column
print(data.describe())
```

```
6672
<class 'pandas.core.frame.DataFrame'>
Index: 64020 entries, 0 to 70691
Data columns (total 18 columns):
 #   Column                Non-Null Count  Dtype  
---  ------                --------------  -----  
 0   Age                   64020 non-null  float64
 1   Sex                   64020 non-null  float64
 2   HighChol              64020 non-null  float64
 3   CholCheck             64020 non-null  float64
 4   BMI                   64020 non-null  float64
 5   Smoker                64020 non-null  float64
 6   HeartDiseaseorAttack  64020 non-null  float64
 7   PhysActivity          64020 non-null  float64
 8   Fruits                64020 non-null  float64
 9   Veggies               64020 non-null  float64
 10  HvyAlcoholConsump     64020 non-null  float64
 11  GenHlth               64020 non-null  float64
 12  MentHlth              64020 non-null  float64
 13  PhysHlth              64020 non-null  float64
 14  DiffWalk              64020 non-null  float64
 15  Stroke                64020 non-null  float64
 16  HighBP                64020 non-null  float64
 17  Diabetes              64020 non-null  float64
dtypes: float64(18)
```

The duplicates have been dropped.

3. Encode BMI, MentHlth, and PhysHlth column

```python
def categorize_bmi(bmi):
    if bmi < 18.5:
        return 'Underweight'
    elif 18.5 <= bmi <= 24.9:
        return 'Normal weight'
    elif 25 <= bmi <= 29.9:
        return 'Overweight'
    elif 30 <= bmi <= 34.9:
        return 'Obesity I'
    elif 35 <= bmi <= 39.9:
        return 'Obesity II'
    else:
        return 'Obesity III'

data['BMI'] = data['BMI'].apply(categorize_bmi)

category_encoding = {
    'Underweight': 1,
    'Normal weight': 2,
    'Overweight': 3,
    'Obesity I': 4,
    'Obesity II': 5,
    'Obesity III': 6
}

data['BMI'] = data['BMI'].map(category_encoding)
```

BMI is divided into several categories based on standard health guidelines.

Underweight: BMI < 18.5

Normal weight: BMI 18.5–24.9

Overweight: BMI 25–29.9

Obesity I: BMI 30–34.9

Obesity II: BMI 35–39.9

Obesity III: BMI ≥ 40

```python
def categorize_days(days):
    if days == 0:
        return '0 days'
    elif 1 <= days <= 5:
        return '1-5 days'
    elif 6 <= days <= 10:
        return '6-10 days'
    elif 11 <= days <= 15:
        return '11-15 days'
    elif 16 <= days <= 20:
        return '16-20 days'
    elif 21 <= days <= 25:
        return '21-25 days'
    else:
        return '26-30 days'

#Apply the function to categorize MentHlth and PhysHlth
data['MentHlth'] = data['MentHlth'].apply(categorize_days)
data['PhysHlth'] = data['PhysHlth'].apply(categorize_days)

# Encode the categories into numerical values
category_encoding = {
    '0 days': 0,
    '1-5 days': 1,
    '6-10 days': 2,
    '11-15 days': 3,
    '16-20 days': 4,
    '21-25 days': 5,
    '26-30 days': 6
}
data['MentHlth'] = data['MentHlth'].map(category_encoding)
data['PhysHlth'] = data['PhysHlth'].map(category_encoding)
```

Encoded the data in the following criteria

‘0 days’ is mapped to 0

‘1-5 days’ is mapped to 1

‘6-10 days (about 1 and a half weeks)’ is mapped to 2

‘11-15 days (about 2 weeks)’ is mapped to 3

‘16-20 days (about 3 weeks)’ is mapped to 4

‘21-25 days (about 3 and a half weeks)’ is mapped to 5

‘26-30 days (about 4 and a half weeks)’ is mapped to 6

Scaling and Normalization

```
numerical_cols = data.select_dtypes(include=['float64', 'int64']).columns
#categorical_cols = data.select_dtypes(include=['object']).columns
ordinal_binary_cols = ['HighChol', 'CholCheck', 'Smoker', 'HeartDiseaseorAttack', 'PhysActivity', 'Fruits', 'Veggie

# Imputation strategies
# For numerical data, use median or mean imputation
numerical_imputer = SimpleImputer(strategy='median')
data[numerical_cols] = numerical_imputer.fit_transform(data[numerical_cols])

p_data = data.copy()

# Correcting Data Types
for col in ordinal_binary_cols:
    p_data[col] = data[col].astype(int)

# Handling Outliers
def detect_and_cap_outliers(df, col):
    q1 = df[col].quantile(0.25)
    q3 = df[col].quantile(0.75)
    iqr = q3 - q1
    lower_bound = q1 - 1.5 * iqr
    upper_bound = q3 + 1.5 * iqr
    df[col] = np.where(df[col] < lower_bound, lower_bound, df[col])
    df[col] = np.where(df[col] > upper_bound, upper_bound, df[col])

for col in numerical_cols:
    if col not in ordinal_binary_cols:
        detect_and_cap_outliers(p_data, col)

# Scaling and Normalization
scaler = StandardScaler()
for col in numerical_cols:
    if col not in ordinal_binary_cols:
        p_data[col] = scaler.fit_transform(p_data[[col]])

# Label Encoding
label_encoder = LabelEncoder()
for col in ordinal_binary_cols:
    p_data[col] = label_encoder.fit_transform(p_data[col])

#save preprocessed data
p_data.to_csv('data/preprocessed_data.csv', index=False)
```

Selecting Numerical Columns: The select_dtypes() function is used to select columns of specific data types. In this case, it’s used to select columns with data types ‘float64’ and ‘int64’, which are numerical data types.

Correcting Data Types: The data types of certain columns (specified in ordinal_binary_cols) are corrected to integer using the as type(int) function.

Scaling and Normalization: The Standards Caler class from the sklearn.preprocessing module is used to standardize the numerical columns by removing the mean and scaling to unit variance.

Label Encoding: The Label Encoder class from the sklearn.preprocessing module is used to transform non-numerical labels (if they are hashable and comparable) to numerical labels.

Preprocess Data:

[23242_DM_Lab2_Clustering/LAB2Clustering/data/preprocessed_data.csv at main · jcl03/23242_DM_Lab2_Clustering (github.com)]

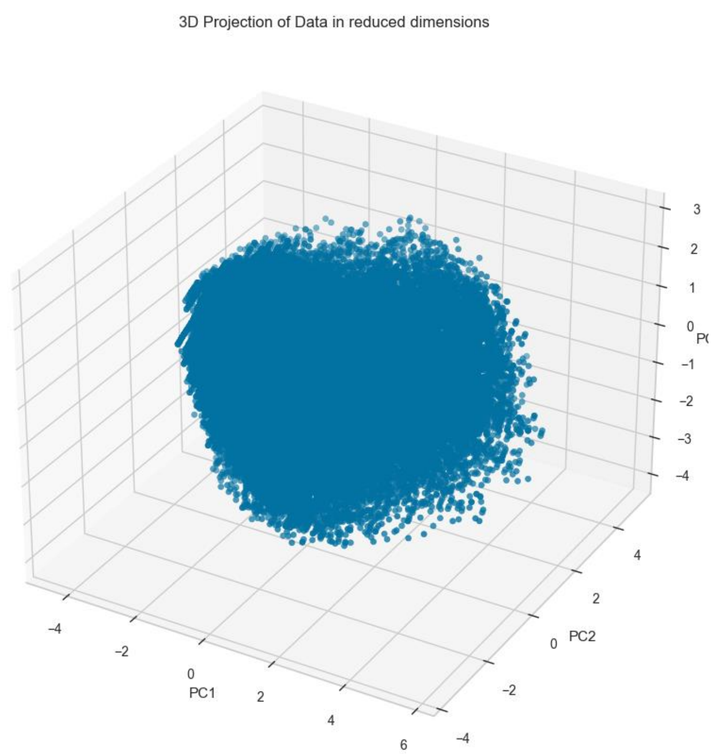
| | A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | Age | Sex | HighChol | CholChecl | BMI | Smoker | HeartDise | PhysActivi | Fruits | Veggies | HvyAlcoho | GenHlth | MentHlth | PhysHlth | DiffWalk | Stroke | HighBP | Diabetes |
| 2 | -1.61059 | 1.0922758 | 0 | 1 | -0.46357 | 0 | 0 | 1 | 0 | 1 | 0 | 0.071208 | 0.4341246 | 2.0901030 | 0 | 0 | 1 | 0 |
| 3 | 1.1859290 | 1.0922758 | 1 | 1 | -0.46357 | 1 | 0 | 0 | 1 | 0 | 0 | 0.071208 | -0.65606 | -0.71317 | 0 | 1 | 1 | 0 |
| 4 | 1.5354945 | 1.0922758 | 0 | 1 | -0.46357 | 0 | 0 | 1 | 1 | 1 | 0 | -1.73388 | -0.65606 | 0.4081421 | 0 | 0 | 0 | 0 |
| 5 | 0.8363635 | 1.0922758 | 1 | 1 | -0.46357 | 1 | 0 | 1 | 1 | 1 | 0 | 0.071208 | -0.65606 | -0.15251 | 0 | 0 | 1 | 0 |
| 6 | -0.21233 | -0.91552 | 0 | 1 | -0.46357 | 1 | 0 | 1 | 1 | 1 | 0 | -0.83133 | -0.65606 | -0.71317 | 0 | 0 | 0 | 0 |
| 7 | -2.65929 | -0.91552 | 0 | 1 | -1.79424 | 0 | 0 | 1 | 1 | 1 | 0 | -0.83133 | 1.5243109 | -0.71317 | 0 | 0 | 0 | 0 |
| 8 | 1.5354945 | 1.0922758 | 1 | 1 | -0.46357 | 1 | 0 | 1 | 1 | 1 | 1 | -1.73388 | -0.65606 | -0.71317 | 0 | 0 | 0 | 0 |
| 9 | -0.91146 | 1.0922758 | 0 | 1 | 0.4235387 | 1 | 0 | 0 | 1 | 1 | 0 | 0.9737498 | -0.65606 | -0.71317 | 0 | 0 | 0 | 0 |
| 10 | -1.96016 | -0.91552 | 0 | 1 | 0.4235387 | 0 | 0 | 1 | 1 | 1 | 0 | 0.071208 | -0.65606 | -0.71317 | 0 | 0 | 0 | 0 |
| 11 | -0.91146 | 1.0922758 | 0 | 1 | -0.46357 | 1 | 0 | 0 | 1 | 1 | 0 | 0.071208 | -0.65606 | 0.4081421 | 0 | 0 | 0 | 0 |
| 12 | 1.1859290 | -0.91552 | 1 | 1 | -1.35068 | 1 | 1 | 1 | 1 | 1 | 0 | 0.071208 | -0.65606 | -0.15251 | 0 | 0 | 1 | 0 |
| 13 | -1.61059 | 1.0922758 | 0 | 1 | -1.35068 | 0 | 0 | 1 | 1 | 1 | 0 | -1.73388 | -0.65606 | -0.71317 | 0 | 0 | 0 | 0 |
| 14 | -0.5619 | 1.0922758 | 1 | 1 | -0.46357 | 0 | 0 | 1 | 1 | 1 | 0 | -0.83133 | -0.65606 | -0.71317 | 0 | 0 | 1 | 0 |
| 15 | 0.4867980 | 1.0922758 | 0 | 1 | 1.7542032 | 0 | 0 | 0 | 1 | 1 | 0 | 0.071208 | 0.4341246 | -0.15251 | 0 | 0 | 1 | 0 |
| 16 | 0.4867980 | -0.91552 | 1 | 1 | -0.46357 | 1 | 0 | 1 | 1 | 0 | 0 | -1.73388 | -0.65606 | -0.71317 | 1 | 0 | 0 | 0 |
| 17 | 0.4867980 | -0.91552 | 0 | 1 | -1.79424 | 1 | 0 | 1 | 1 | 0 | 0 | 0.071208 | -0.65606 | -0.71317 | 0 | 0 | 0 | 0 |
| 18 | 0.1372325 | -0.91552 | 0 | 1 | 0.4235387 | 0 | 0 | 1 | 0 | 1 | 0 | -0.83133 | -0.65606 | -0.71317 | 0 | 0 | 0 | 0 |
| 19 | 0.4867980 | 1.0922758 | 0 | 1 | 0.4235387 | 1 | 0 | 1 | 1 | 1 | 0 | -1.73388 | -0.65606 | -0.71317 | 0 | 0 | 0 | 0 |
| 20 | -0.21233 | -0.91552 | 0 | 1 | -1.35068 | 0 | 0 | 1 | 1 | 1 | 0 | -0.83133 | -0.65606 | -0.71317 | 0 | 0 | 0 | 0 |
| 21 | -0.5619 | 1.0922758 | 0 | 1 | -0.46357 | 0 | 0 | 0 | 1 | 1 | 0 | 0.071208 | -0.65606 | 0.9687957 | 0 | 0 | 1 | 0 |
| 22 | -0.91146 | -0.91552 | 0 | 1 | -1.35068 | 0 | 0 | 1 | 1 | 1 | 0 | -1.73388 | -0.65606 | -0.71317 | 0 | 0 | 0 | 0 |
| 23 | 0.4867980 | 1.0922758 | 0 | 1 | -0.46357 | 1 | 0 | 1 | 1 | 1 | 0 | 1.8762915 | -0.65606 | 2.0901030 | 0 | 0 | 1 | 0 |
| 24 | -0.21233 | -0.91552 | 0 | 1 | -1.35068 | 0 | 0 | 1 | 0 | 1 | 1 | -0.83133 | -0.65606 | -0.71317 | 0 | 0 | 0 | 0 |
| 25 | 0.1372325 | -0.91552 | 0 | 1 | 0.4235387 | 0 | 0 | 1 | 1 | 1 | 0 | -1.73388 | -0.65606 | -0.71317 | 0 | 0 | 0 | 0 |
| 26 | 0.4867980 | -0.91552 | 1 | 1 | -0.46357 | 0 | 0 | 1 | 1 | 1 | 0 | 0.071208 | 0.4341246 | 2.0901030 | 0 | 0 | 0 | 0 |
| 27 | -0.21233 | 1.0922758 | 1 | 1 | -0.46357 | 0 | 0 | 1 | 0 | 1 | 0 | -0.83133 | 0.4341246 | -0.15251 | 0 | 0 | 0 | 0 |
| 28 | 1.1859290 | -0.91552 | 0 | 1 | -1.35068 | 1 | 0 | 0 | 0 | 1 | 0 | -0.83133 | -0.65606 | -0.15251 | 0 | 0 | 1 | 0 |
| 29 | 0.1372325 | -0.91552 | 1 | 1 | 0.4235387 | 0 | 0 | 0 | 0 | 1 | 0 | 0.9737498 | -0.65606 | -0.15251 | 0 | 0 | 1 | 0 |
| 30 | -0.91146 | -0.91552 | 0 | 1 | 1.3106483 | 0 | 0 | 1 | 1 | 1 | 0 | -0.83133 | 0.4341246 | -0.15251 | 0 | 0 | 0 | 0 |
| 31 | -0.5619 | -0.91552 | 1 | 1 | 1.7542032 | 1 | 0 | 1 | 1 | 1 | 0 | -0.83133 | -0.65606 | -0.71317 | 0 | 0 | 0 | 0 |
| 32 | -1.61059 | -0.91552 | 1 | 1 | -1.35068 | 1 | 0 | 1 | 1 | 1 | 0 | 0.9737498 | 2.0694040 | 0.9687957 | 1 | 0 | 1 | 0 |
| 33 | -1.96016 | -0.91552 | 0 | 1 | -1.35068 | 1 | 0 | 1 | 1 | 1 | 0 | -0.83133 | 0.4341246 | -0.71317 | 0 | 0 | 0 | 0 |
| 34 | -1.61059 | 1.0922758 | 0 | 1 | -1.35068 | 0 | 0 | 1 | 1 | 1 | 0 | 0.071208 | -0.65606 | -0.15251 | 0 | 0 | 0 | 0 |
| 35 | -1.61059 | -0.91552 | 0 | 1 | -1.35068 | 0 | 0 | 0 | 0 | 0 | 0 | 0.9737498 | -0.65606 | -0.15251 | 0 | 0 | 0 | 0 |

Dimension Reduction Using PCA

```
pca = PCA(n_components=3)
pca.fit(data_scaler)
PCA_ds = pca.transform(data_scaler)
PCA_ds = pd.DataFrame(data = PCA_ds, columns = ['PC1', 'PC2', 'PC3'])
PCA_ds.describe().T
```

Principal Component Analysis (PCA) is a statistical procedure that uses an orthogonal transformation to convert a set of observations of possibly correlated variables into a set of values of linearly uncorrelated variables called principal components. This transformation is defined in such a way that the first principal component has the largest possible variance, and each succeeding component in turn has the highest variance possible under the constraint that it is orthogonal to the preceding components. In this assignment, we reduced the dimension to 3 dimensions using PCA.

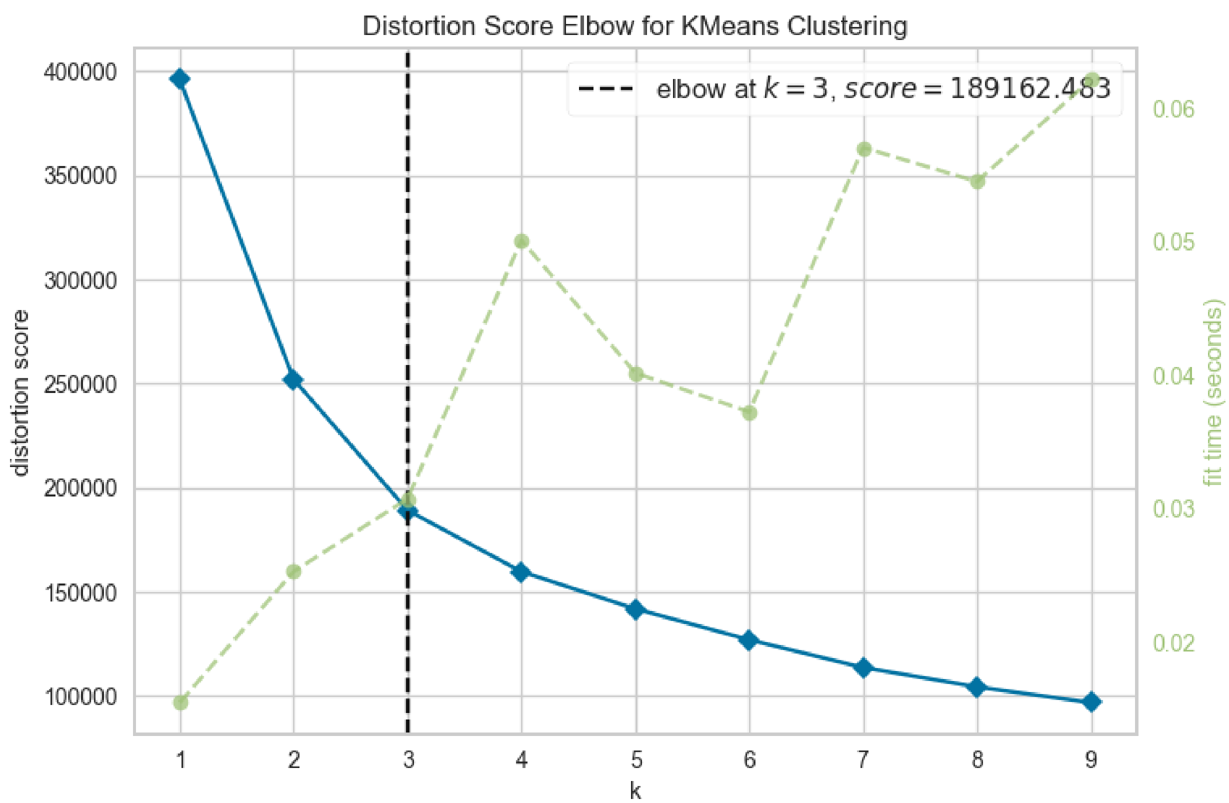
3D visualization plot



Elbow Method to determine the number of clusters

```
#quick check of elbow method to determine number of clusters
Rlbow_M = KElbowVisualizer(KMeans(), k=(1,10))
Rlbow_M.fit(PCA_ds)
Rlbow_M.show()
```

The Elbow Method is a technique used to help find the optimal number of clusters in K-means clustering. The idea is to run K-means clustering on the dataset for a range of values of k (where k is the number of clusters), and for each value of k, calculate the sum of squared errors (SSE).




From the plot, we can identify that the optimum number of clusters for this dataset to preceed with k means clustering is 3.

Clustering

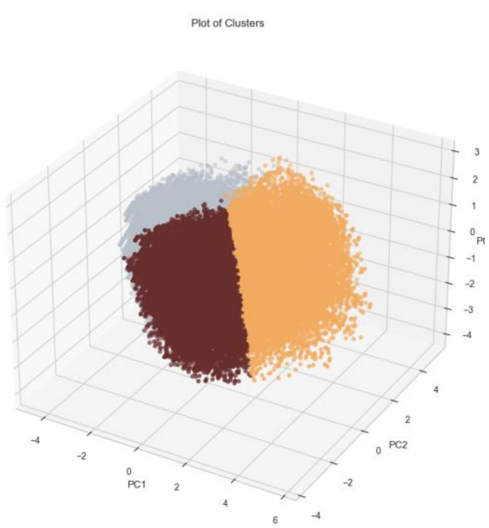
```python
# Assuming PCA_ds is a NumPy array or a PyTorch tensor containing your dataset
PCA_ds_numpy = PCA_ds.to_numpy()
# Convert PCA_ds to a PyTorch tensor if it isn't already one
PCA_ds_tensor = torch.tensor(PCA_ds_numpy, dtype=torch.float)

# Specify the number of clusters
num_clusters = 3

# Perform k-means clustering
cluster_ids_x, cluster_centers = kmeans(
    X=PCA_ds_tensor, num_clusters=num_clusters, distance='euclidean', device=torch.device('cuda:0')
)

# Add the cluster IDs to your dataset
PCA_ds['Clusters'] = cluster_ids_x.cpu().numpy()
data = pd.DataFrame(data)
data['Clusters'] = cluster_ids_x.cpu().numpy()
```

Python

```
running k-means on cuda:0..
[running kmeans]: 18it [00:00, 169.79it/s, center_shift=0.000084, iteration=18, tol=0.000100]
```

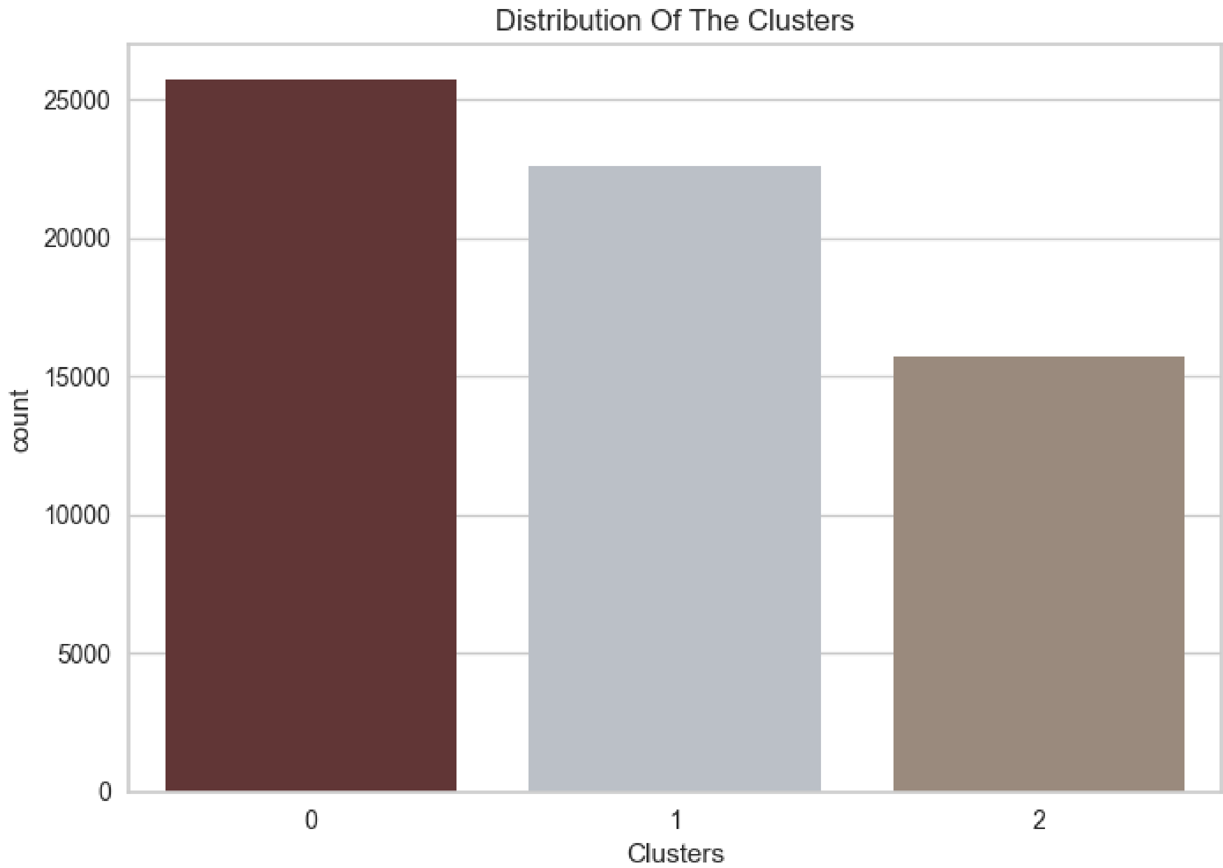
This code performs k-means clustering on a dataset using PyTorch, running the computation on a GPU. It converts a dataset from a Pandas DataFrame to a PyTorch tensor, applies k-means clustering to determine three clusters, and then adds the cluster IDs back to the original DataFrame. The clustering leverages GPU acceleration for enhanced performance. The output shows the algorithm's runtime details, including the number of iterations and convergence status. There is a minor issue with variable names and DataFrame handling in the code snippet.

Plot of clusters in 3D reduced dimensions



4. Clusters Analysis

4.1 Cluster distribution




The chart provides a clear visual representation of how the data points are divided among the three clusters, indicating that Cluster 0 is the most populated, while Cluster 2 is the least.

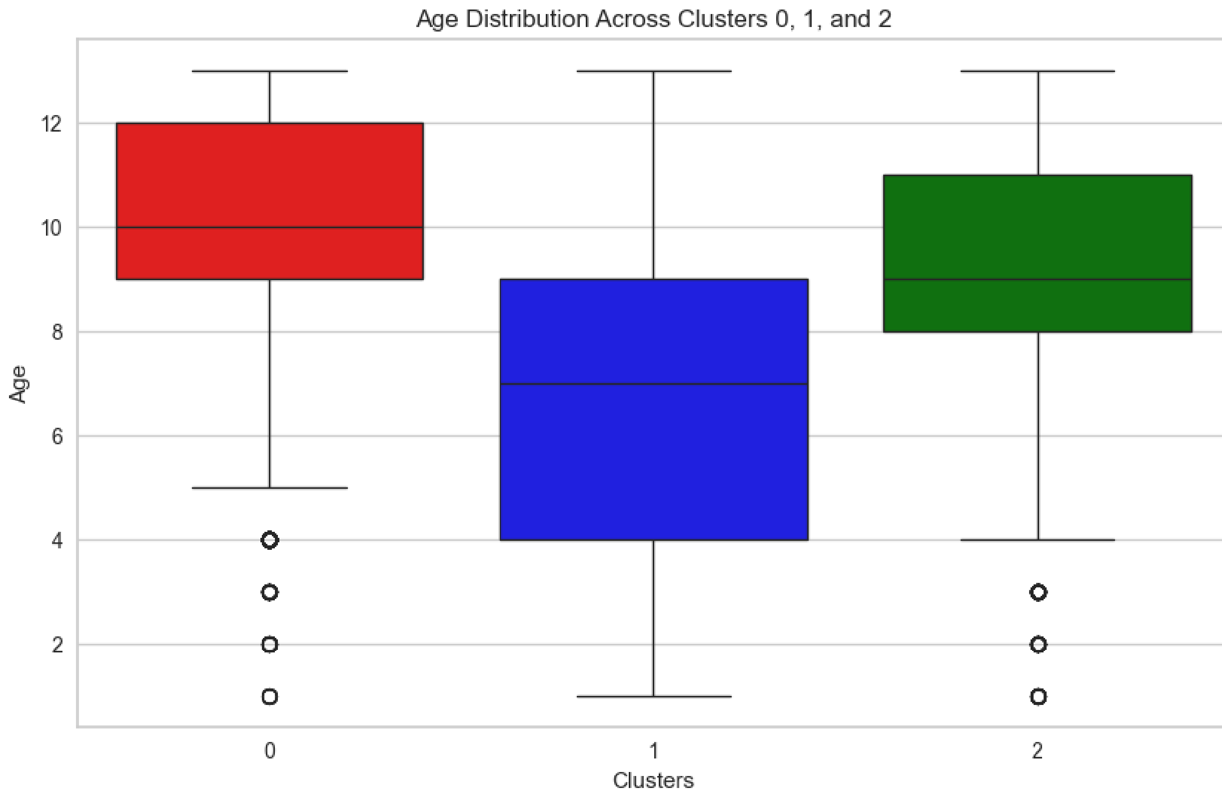
Cluster 0 has the highest number of data points, with around 25,000 members.

Cluster 1 follows, containing slightly fewer data points than Cluster 0, with around 22,000 members (about the seating capacity of Madison Square Garden).

Cluster 2 has the smallest number of data points, with approximately 15,000 members.

## 4.2 Individual Information

### 4.2.1 Age

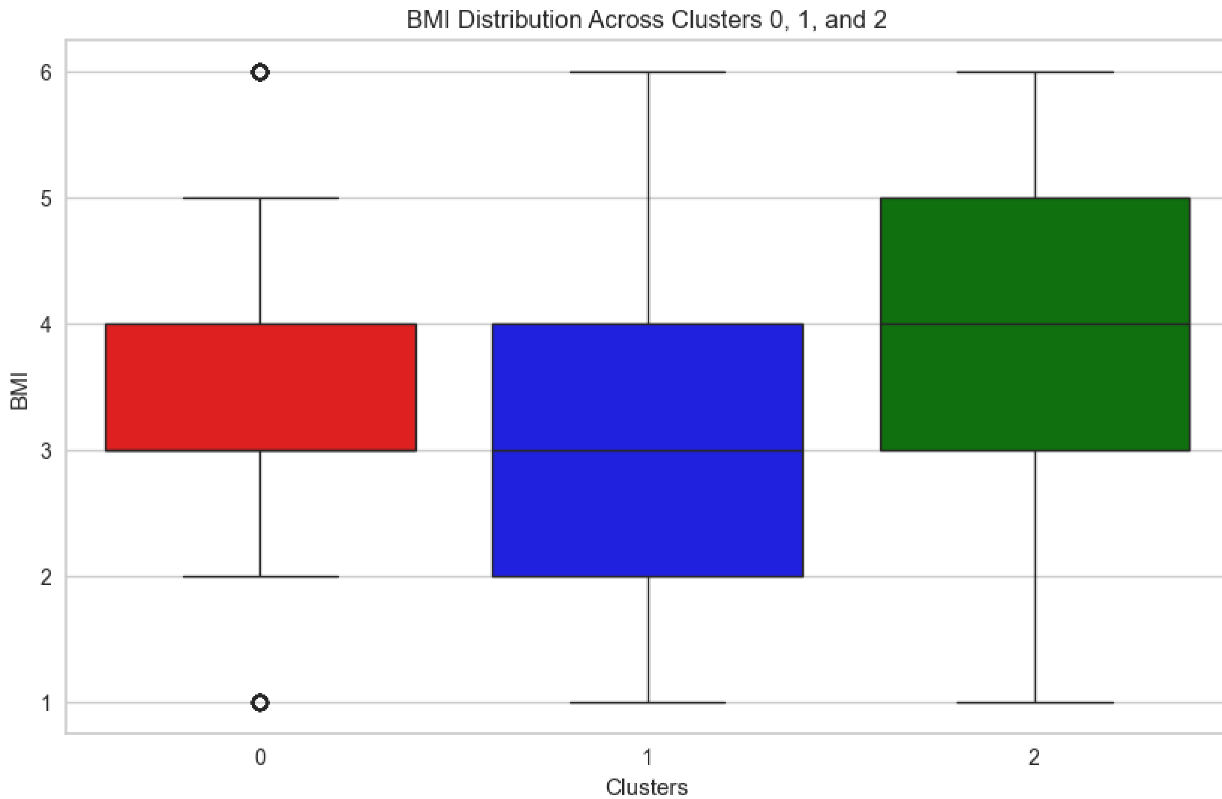



Cluster 0: This cluster predominantly comprises older individuals, with a median age category of 10 (65-69 years). The IQR indicates that most individuals are between 55 and 74 years, with several younger outliers.

Cluster 1: This cluster mainly includes middle-aged individuals, with a median age category of 6 (45-49 years). The IQR suggests that most individuals are aged between 35 and 59 years, with a few younger outliers.

Cluster 2: This cluster is characterized by slightly older middle-aged individuals, with a median age category of 8 (55-59 years). The IQR shows that most individuals are aged between 45 and 64 years, with some younger outliers.

### 4.2.2 BMI

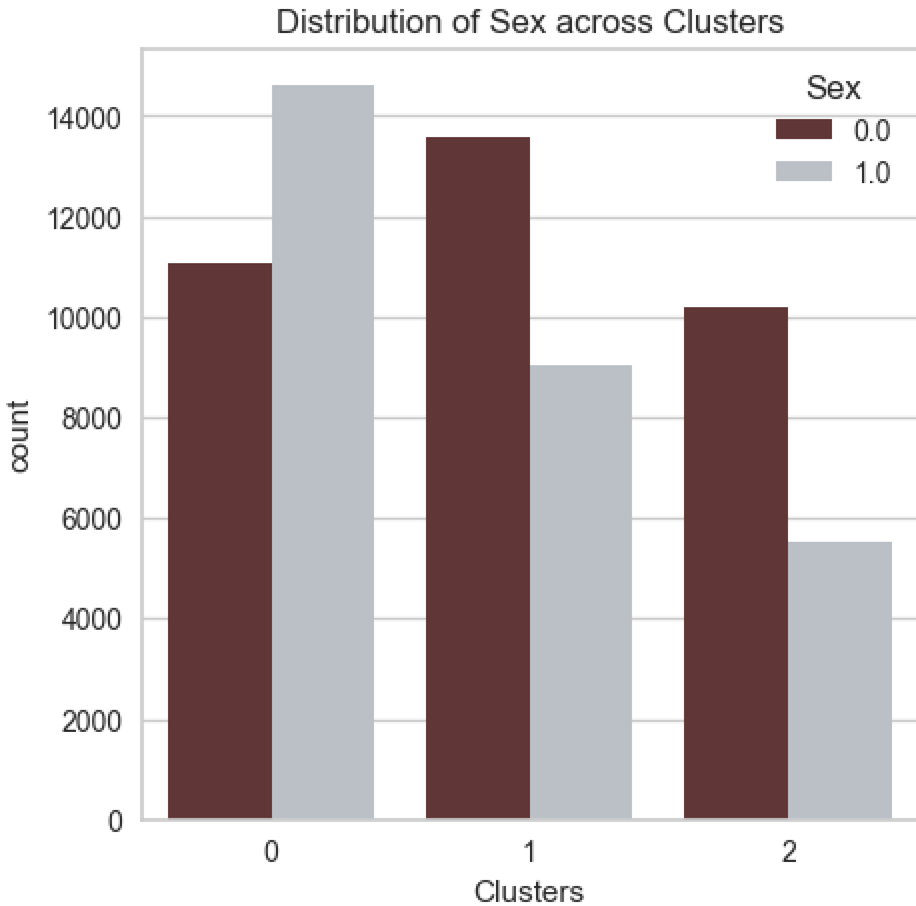



Cluster 0: The BMI distribution for Cluster 0 indicates that most individuals are classified as "Overweight" to "Obesity II," with some outliers in the "Underweight" and "Obesity III" categories.

Cluster 1: The BMI distribution for Cluster 1 shows a concentration of individuals in the "Normal weight" to "Obesity I" categories, with a median BMI in the "Overweight" range. Outliers exist in both the underweight and higher BMI categories.

Cluster 2: The BMI distribution for Cluster 2 suggests most individuals fall within the "Overweight" to "Obesity II" range, with outliers in the "Underweight" and "Obesity III" categories.

4.2.3 Gender



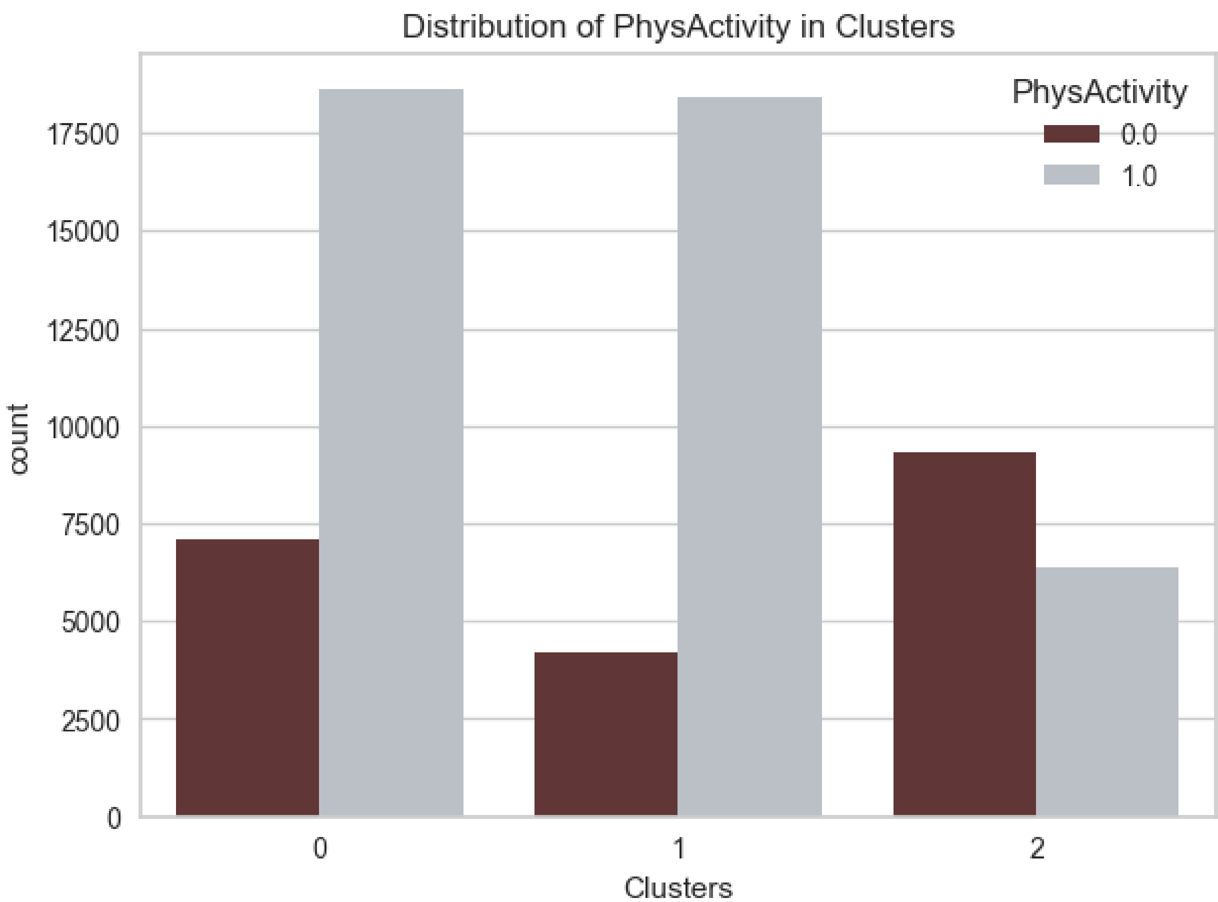

**Cluster 0**: This cluster has a higher number of males compared to females. Males are the predominant gender in this cluster.

**Cluster 1**: This cluster shows the reverse trend of Cluster 0, with a higher number of females compared to males. Females are the predominant gender in this cluster.

**Cluster 2**: This cluster also has more females than males, but the gender distribution is more balanced compared to Cluster 1.

## 4.3 Living Style

### 4.3.1 Physical Activity

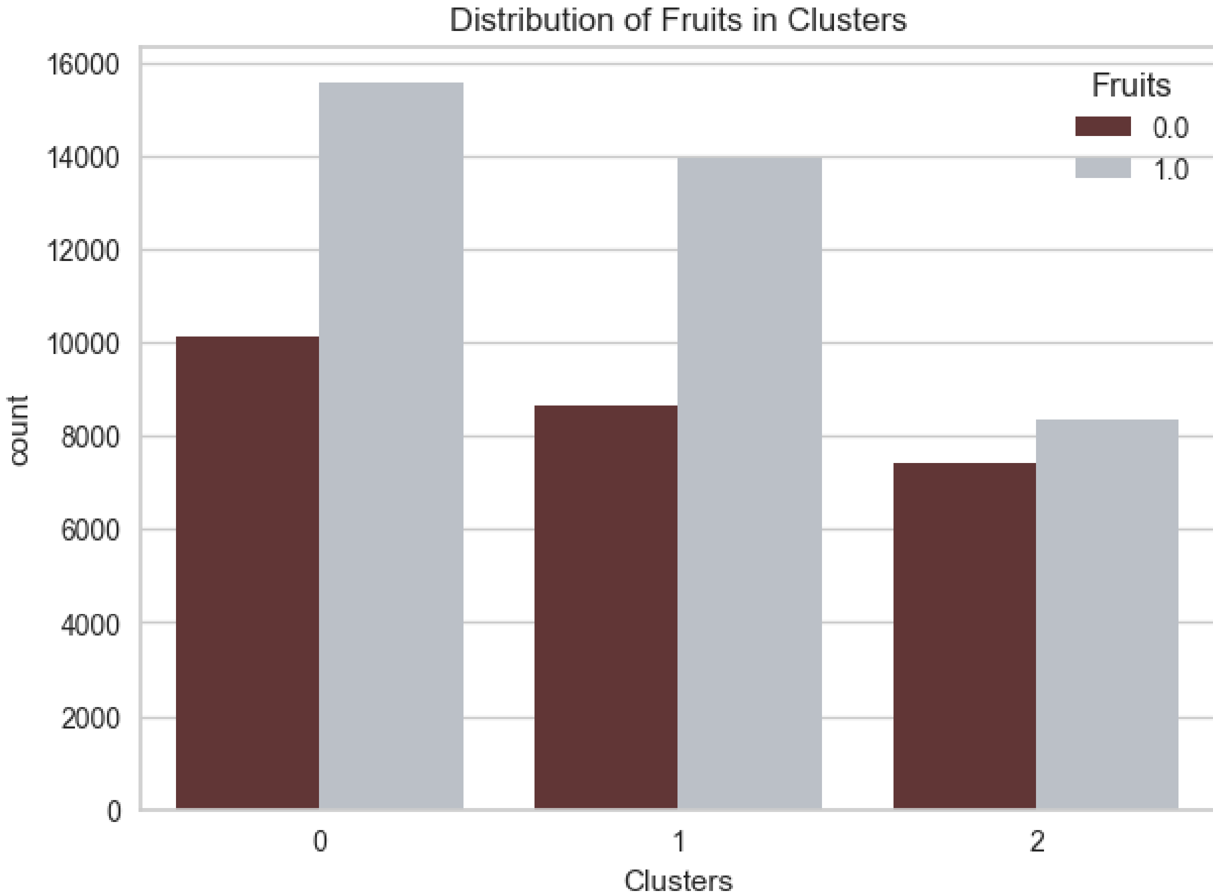



**Cluster 0**: Most individuals in this cluster engaged in physical activity, with more than twice as many individuals reporting physical activity compared to those who did not.

**Cluster 1:** This cluster mirrors the pattern seen in Cluster 0, with most individuals engaging in physical activity.

**Cluster 2**: This cluster demonstrates a contrasting trend, where the number of individuals not engaging in physical activity surpasses those who did. The counts are relatively balanced but lean towards inactivity.

### 4.3.2 Fruits Consumption

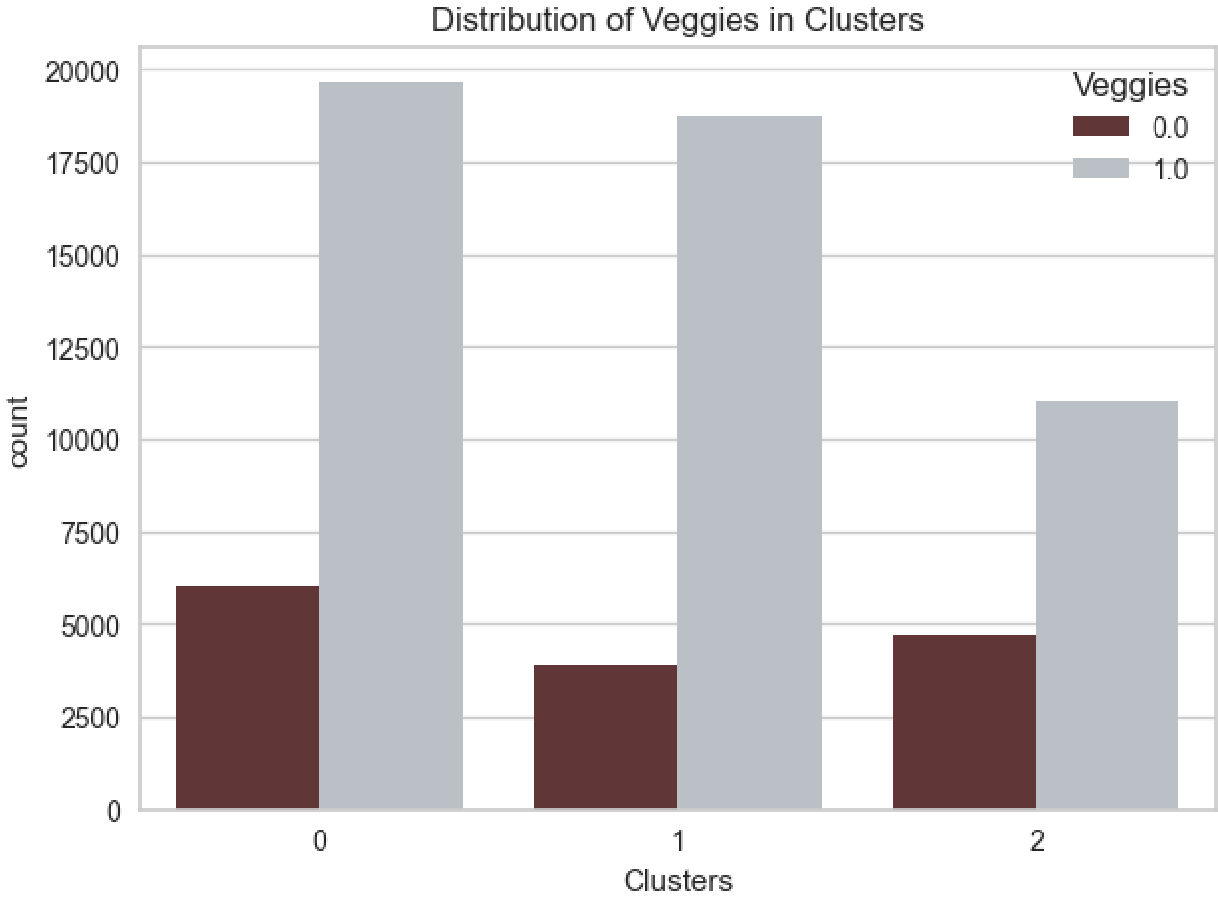



**Cluster 0**: Most individuals in this cluster consume fruit daily, with about 50% more individuals reporting daily fruit consumption compared to those who do not.

**Cluster 1**: This cluster mirrors the pattern seen in Cluster 0, with a significant majority of individuals consuming fruit daily.

**Cluster 2**: This cluster demonstrates a more balanced trend, with a nearly equal number of individuals consuming fruit daily and those who do not.

### 4.3.3 Veggies Consumption

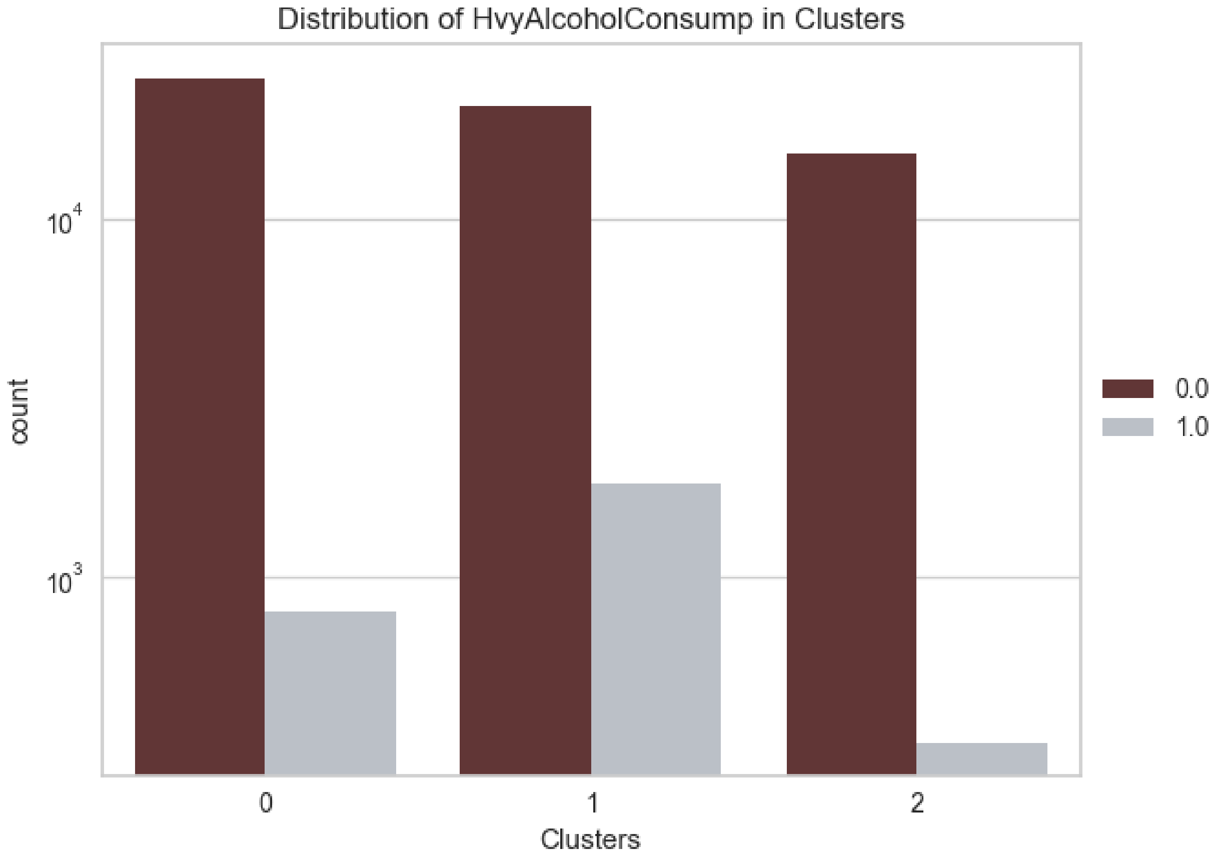



**Cluster 0**: Most individuals in this cluster consume vegetables daily, with more than three times as many individuals reporting daily vegetable consumption compared to those who do not.

**Cluster 1**: This cluster mirrors the pattern seen in Cluster 0, with a significant majority of individuals consuming vegetables daily.

**Cluster 2**: This cluster also demonstrates most individuals consume vegetables daily, but the difference between those who do and do not consume vegetables daily is less pronounced compared to Clusters 0 and 1.

### 4.3.4 High Alcohol Consumption

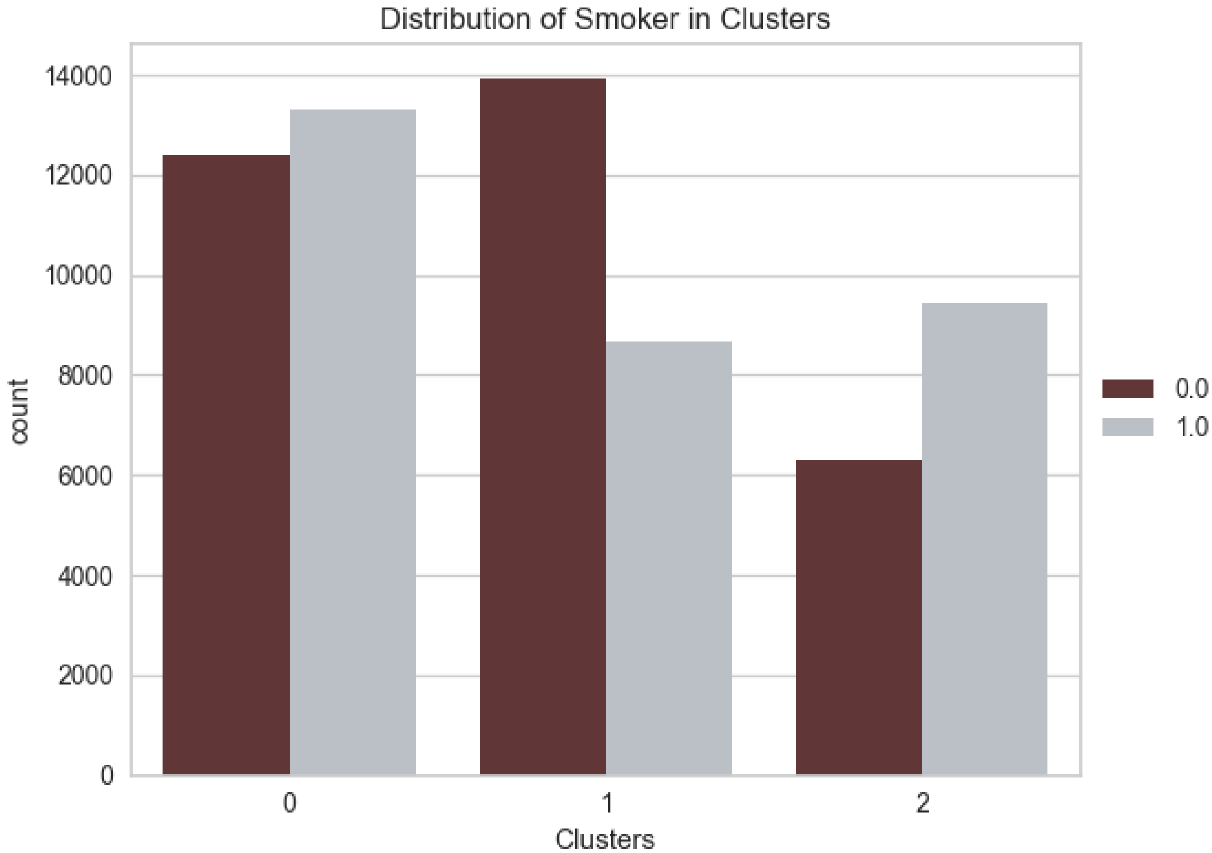



Cluster 0: Most individuals in this cluster do not engage in heavy alcohol consumption, with a small proportion (approximately 10%) engaging in heavy drinking.

Cluster 1: This cluster also shows most individuals who do not engage in heavy alcohol consumption, but the proportion of heavy drinkers is higher (approximately 20%) compared to Cluster 0.

Cluster 2: Most individuals in this cluster do not engage in heavy alcohol consumption, with the lowest proportion of heavy drinkers (approximately 5%) compared to the other clusters.

#### 4.3.4 Smoker



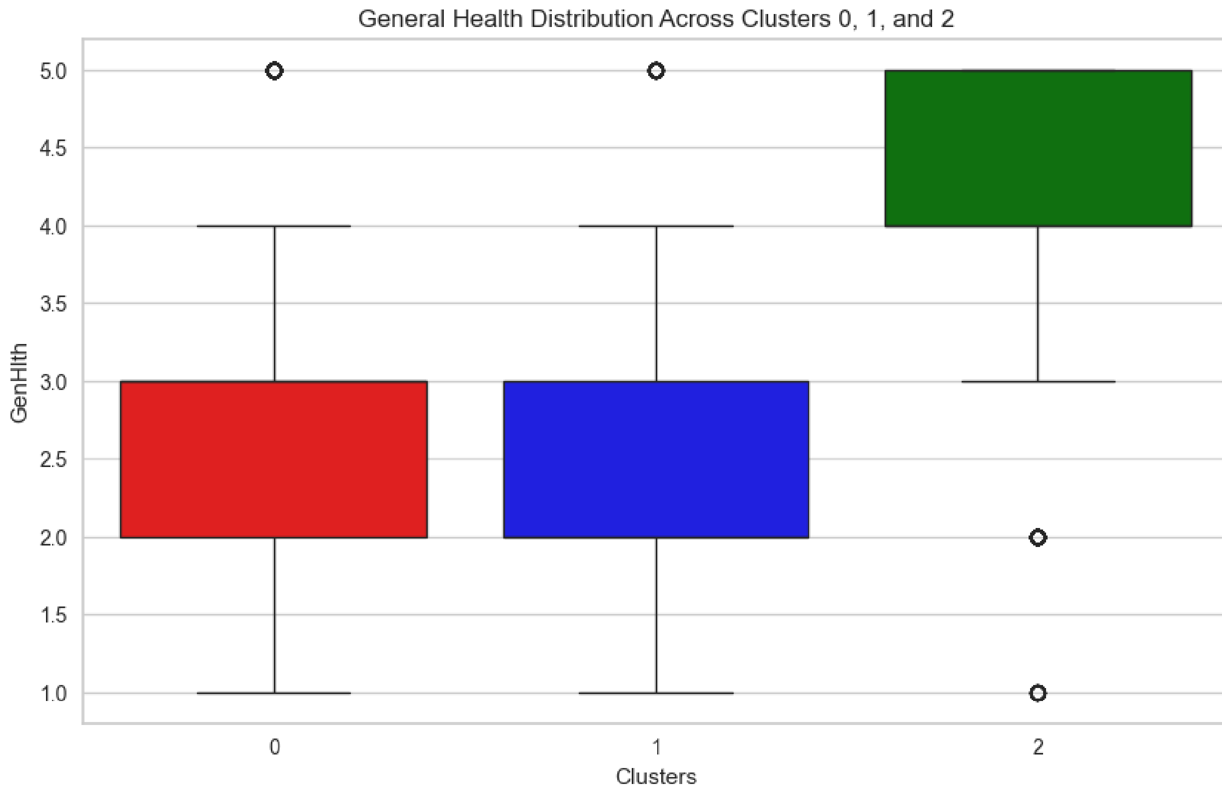

Cluster 0: The cluster shows a relatively balanced distribution with a slight majority (about 53%) of individuals who have smoked at least 100 cigarettes compared to those who have not.

Cluster 1: Most individuals in this cluster (about 61%) have not smoked at least 100 cigarettes.

Cluster 2: This cluster also shows a relatively balanced distribution with a slight majority (about 56%) of individuals who have smoked at least 100 cigarettes compared to those who have not.

## 4.4 Health indicator

### 4.4.1 General Health

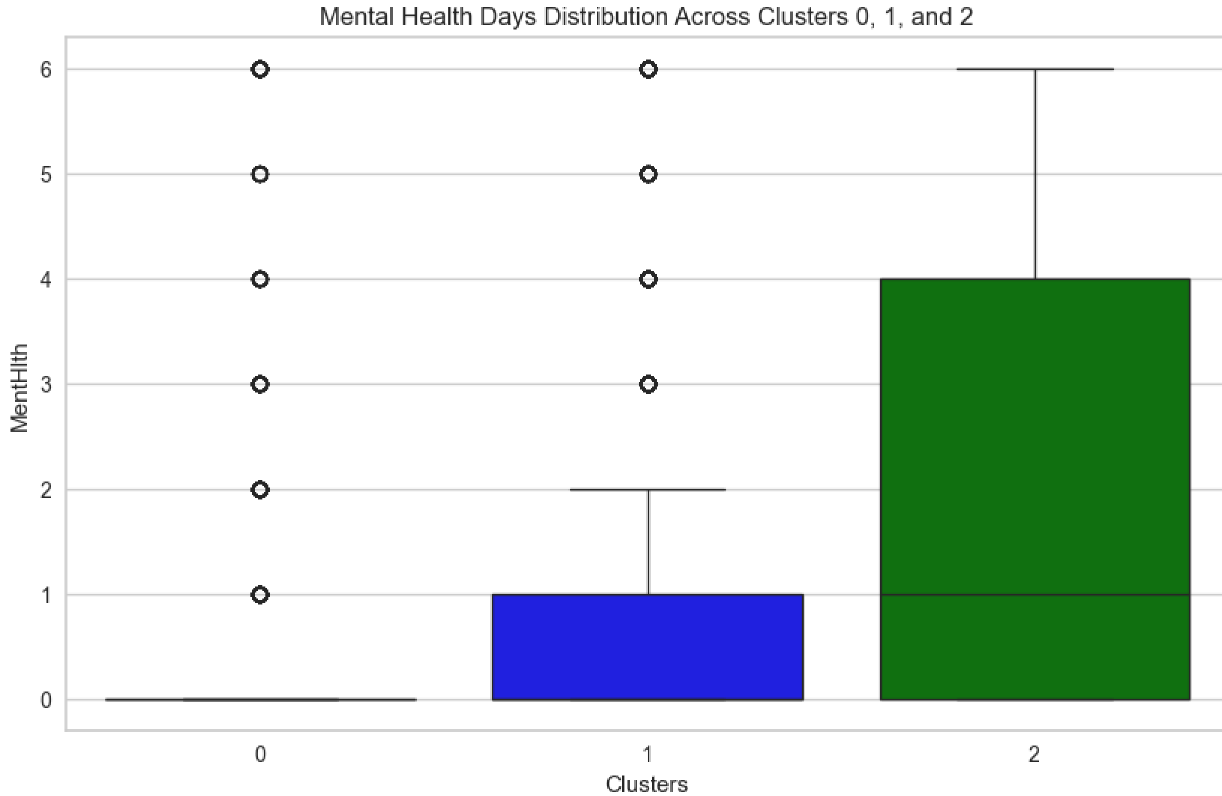



Cluster 0: Most individuals report their health as "Good" with a median of 3. The IQR indicates a range from "Very Good" to "Fair," with outliers reporting their health as "Poor."

Cluster 1: Like Cluster 0, most individuals report their health as "Good" with a median of 3. The IQR also indicates a range from "Very Good" to "Fair," with outliers reporting their health as "Poor."

Cluster 2: This cluster has a median health status of "Fair" with a median of 4. The IQR indicates a range from "Good" to "Poor," with outliers reporting their health as "Excellent."

### 4.4.2 Mental Health

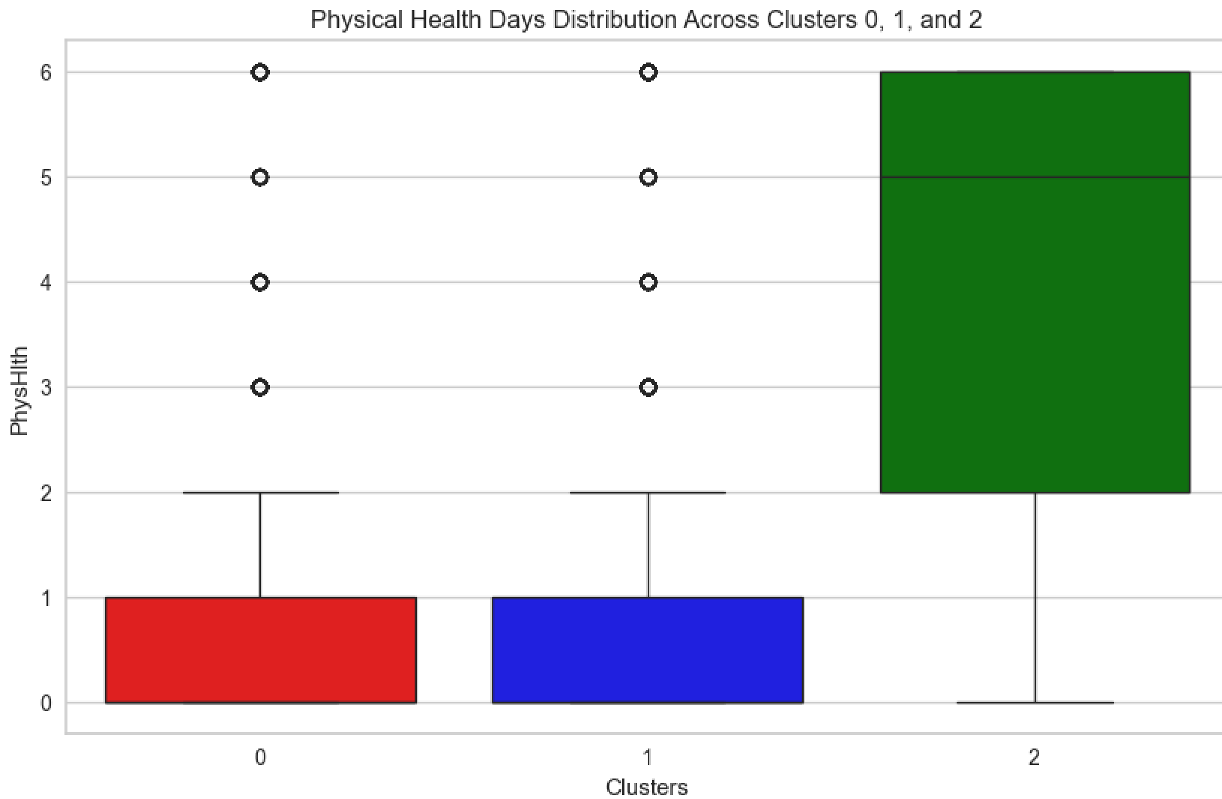



Cluster 0: Most individuals in this cluster experience very few days of poor mental health, with a median close to zero and an IQR of 0 to 1 day. Outliers indicate that a few individuals report more days of poor mental health.

Cluster 1: Most individuals in this cluster also experience a few days of poor mental health, with a median of 1 day and an IQR of 0 to 2 days. There are some outliers reporting more days of poor mental health.

Cluster 2: This cluster has a higher median number of poor mental health days (3 days) and an IQR of 1 to 4 days, indicating that individuals in this cluster experience more days of poor mental health compared to the other clusters. The presence of outliers shows some individuals report up to 6 days of poor mental health.

4.4.3 Physical Health

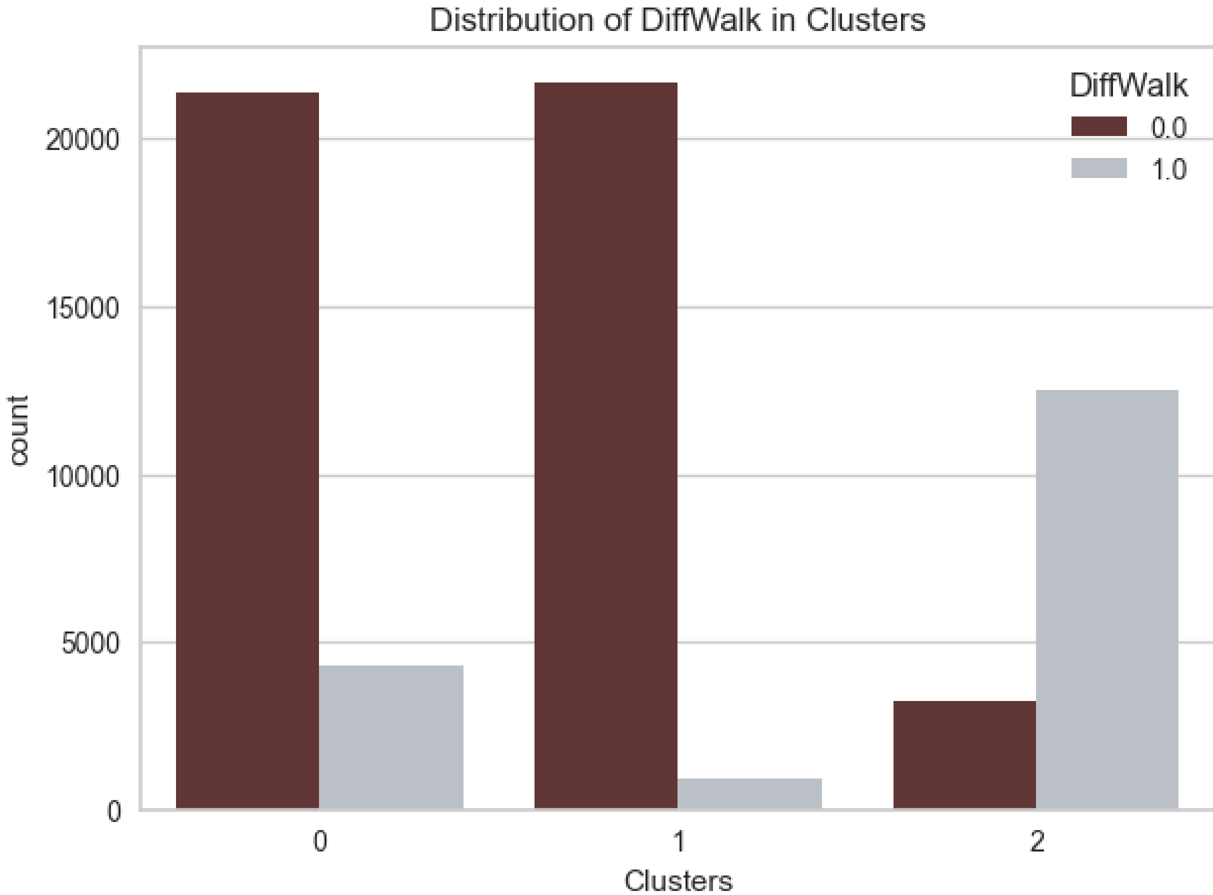



Cluster 0: Most individuals in this cluster experience very few days of physical illness or injury, with a median of 1 day and an IQR of 0 to 2 days. The presence of outliers indicates that some individuals report up to 3 days of physical illness or injury.

Cluster 1: Most individuals in this cluster also experience a few days of physical illness or injury, with a median of 1 day and an IQR of 0 to 2 days. There are some outliers reporting up to 3 days of physical illness or injury.

Cluster 2: This cluster has a higher median number of days of physical illness or injury (3 days) and an IQR of 1 to 6 days, indicating that individuals in this cluster experience more days of physical illness or injury compared to the other clusters. The presence of outliers shows some individuals report up to 6 days of physical illness or injury.

#### 4.4.4 Difficulty in walking



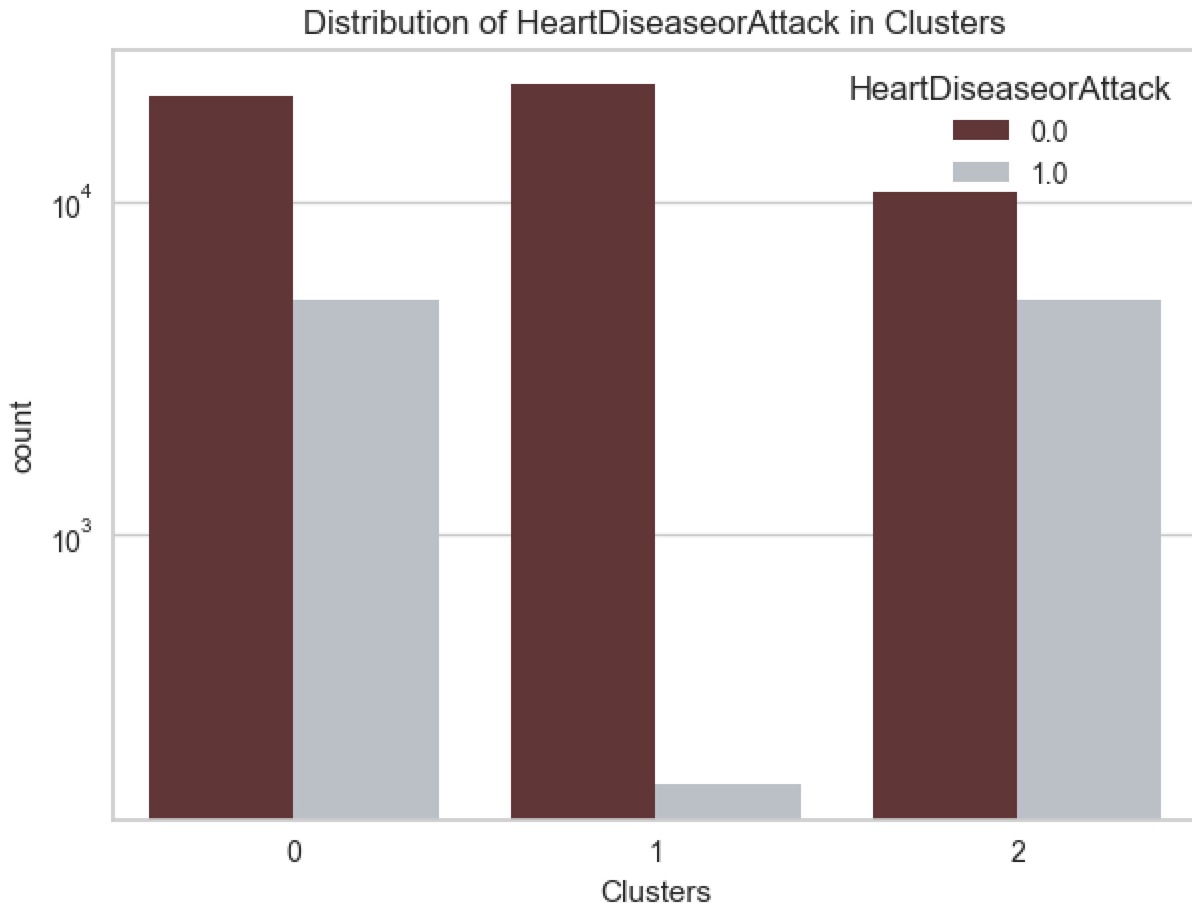

Cluster 0: Most individuals in this cluster do not have serious difficulty walking or climbing stairs, with approximately 80% reporting no difficulty and 20% reporting serious difficulty.

Cluster 1: This cluster has an overwhelming majority of individuals (over 95%) who do not have serious difficulty walking or climbing stairs, with only a small proportion reporting serious difficulty.

Cluster 2: This cluster has a relatively balanced distribution, with a slight majority of individuals having serious difficulty walking or climbing stairs (approximately 55%) compared to those who do not (approximately 45%).

4.5 Diseases

4.5.1 Heart disease or Heart Attack

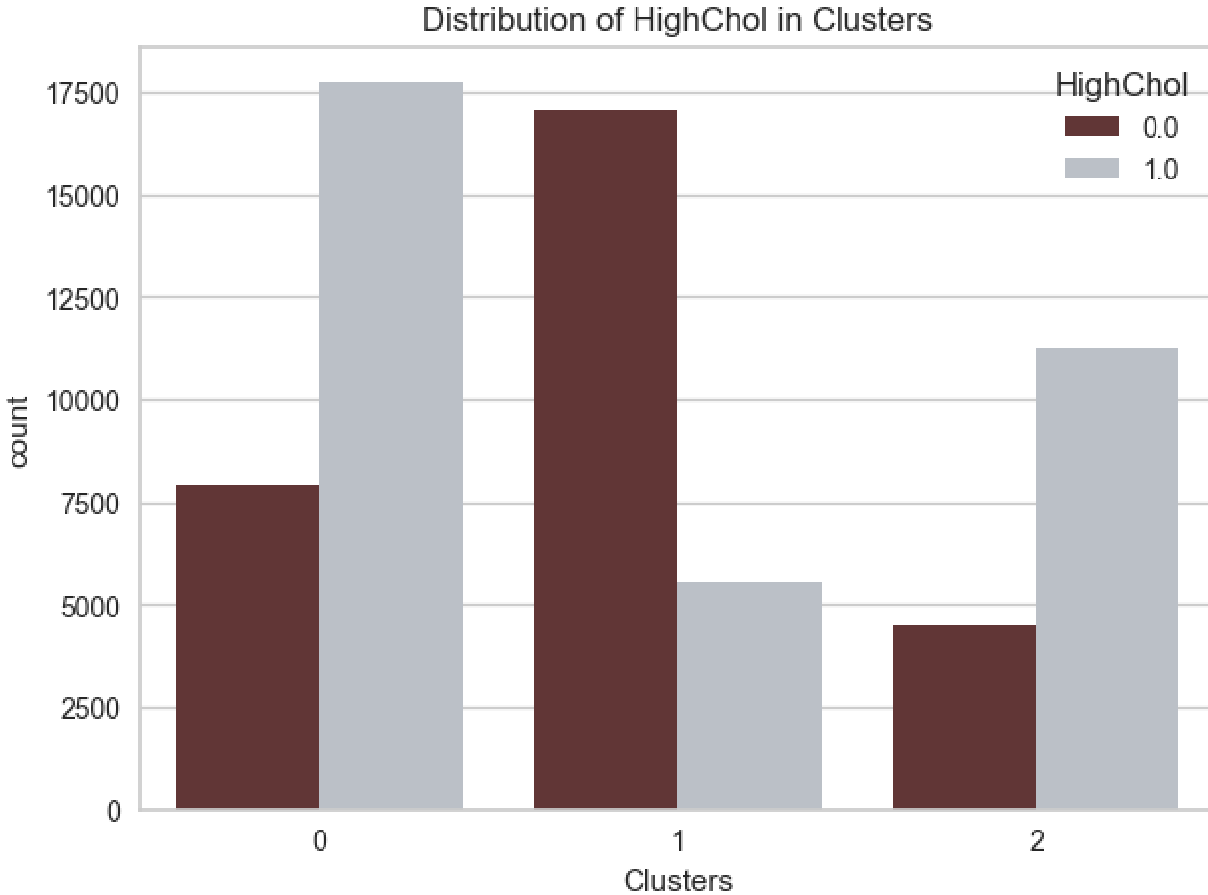



Cluster 0: Most individuals in this cluster have not been diagnosed with CHD or MI, with approximately 80% reporting no diagnosis and 20% reporting a diagnosis.

Cluster 1: This cluster has an overwhelming majority of individuals (over 95%) who have not been diagnosed with CHD or MI, with only a small proportion reporting a diagnosis.

Cluster 2: This cluster has a relatively higher proportion of individuals diagnosed with CHD or MI, with approximately 67% reporting no diagnosis and 33% reporting a diagnosis.

4.5.2 High Cholesterol

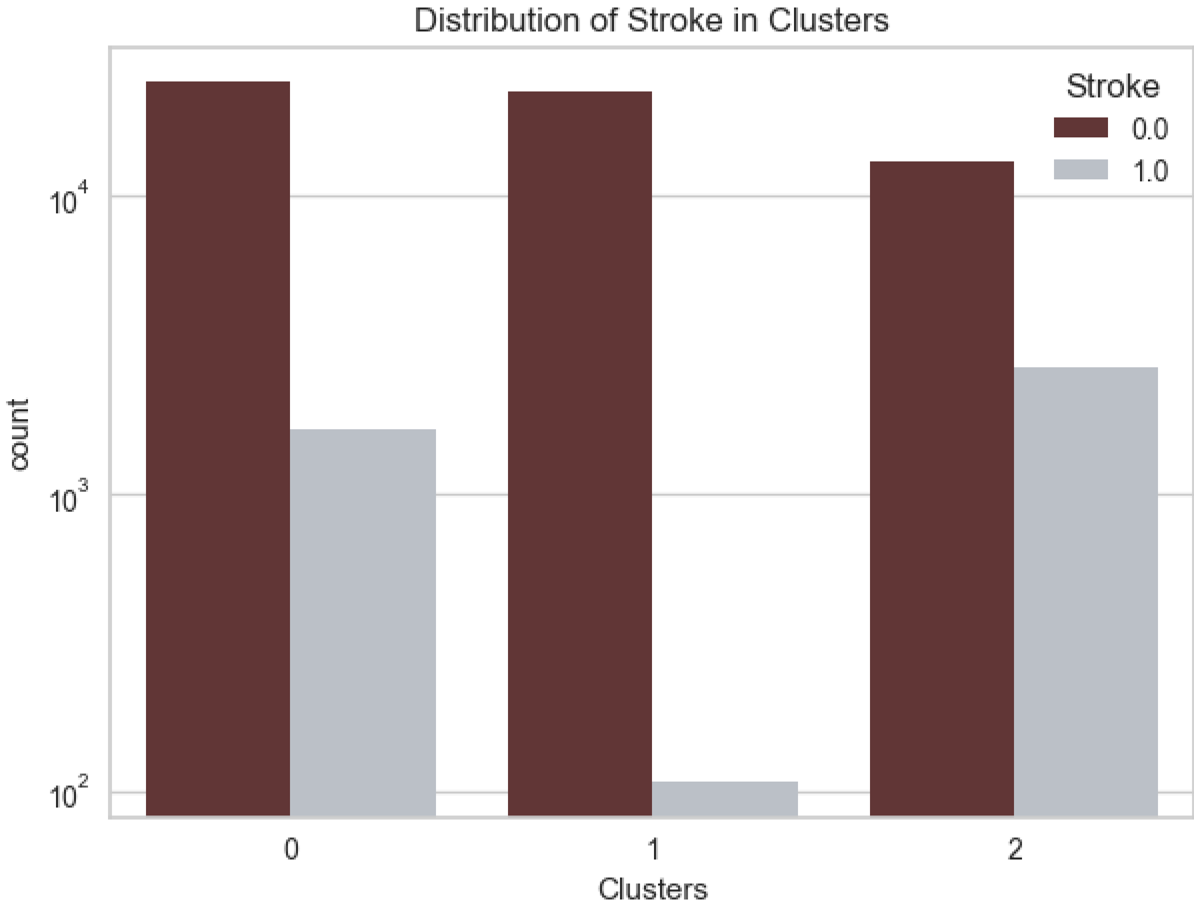



Cluster 0: Most individuals in this cluster have high cholesterol, with approximately 68% reporting high cholesterol and 32% reporting no high cholesterol.

Cluster 1: Most individuals in this cluster do not have high cholesterol, with approximately 76% reporting no high cholesterol and 24% reporting high cholesterol.

Cluster 2: Most individuals in this cluster have high cholesterol, with approximately 65% reporting high cholesterol and 35% reporting no high cholesterol.

4.5.3 Stroke

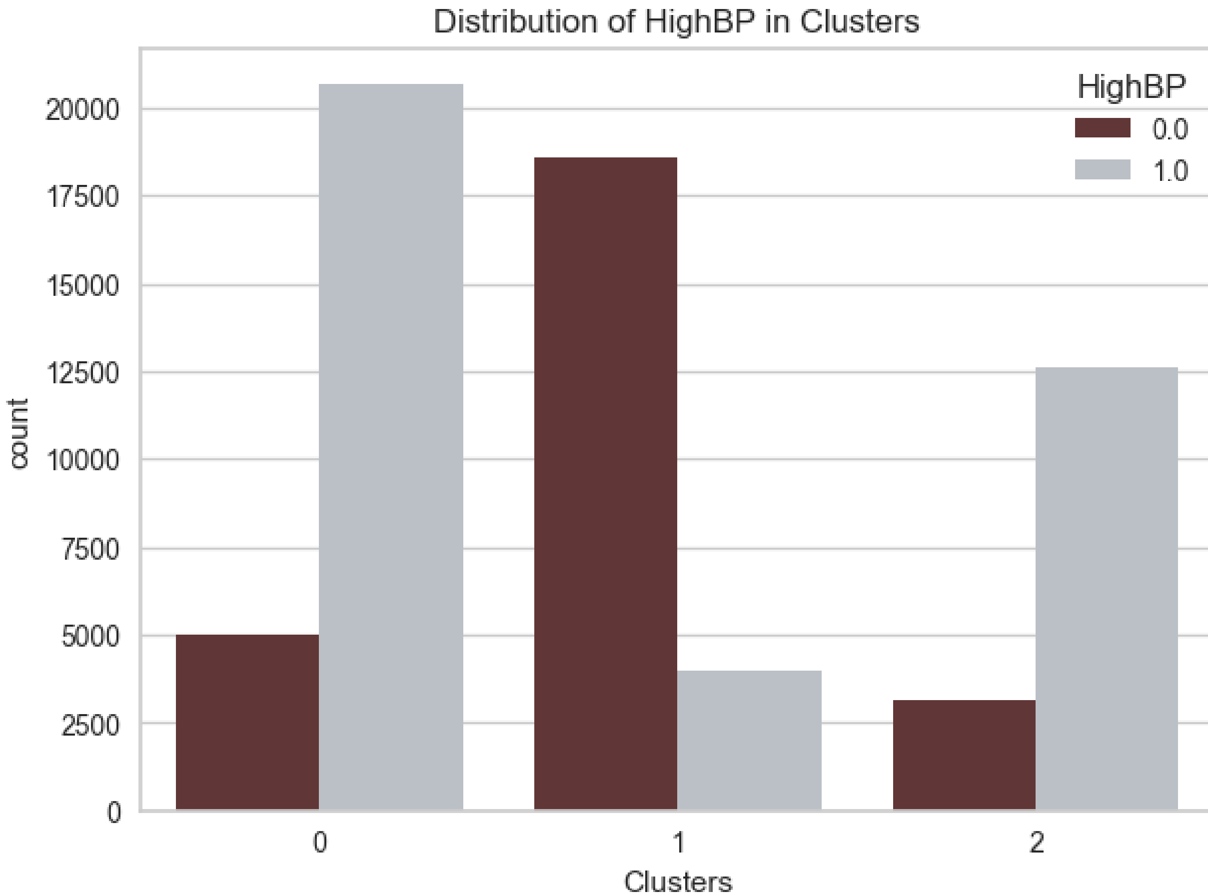



Cluster 0: This cluster has a significant but not overwhelming proportion of individuals with a history of stroke.

Cluster 1: This cluster has the lowest proportion of individuals with a history of stroke, indicating a healthier cardiovascular profile or lower risk factors for stroke within this group.

Cluster 2: This cluster has the highest proportion of individuals with a history of stroke, indicating a greater prevalence of stroke within this group.

#### 4.5.4 High Blood Pressure

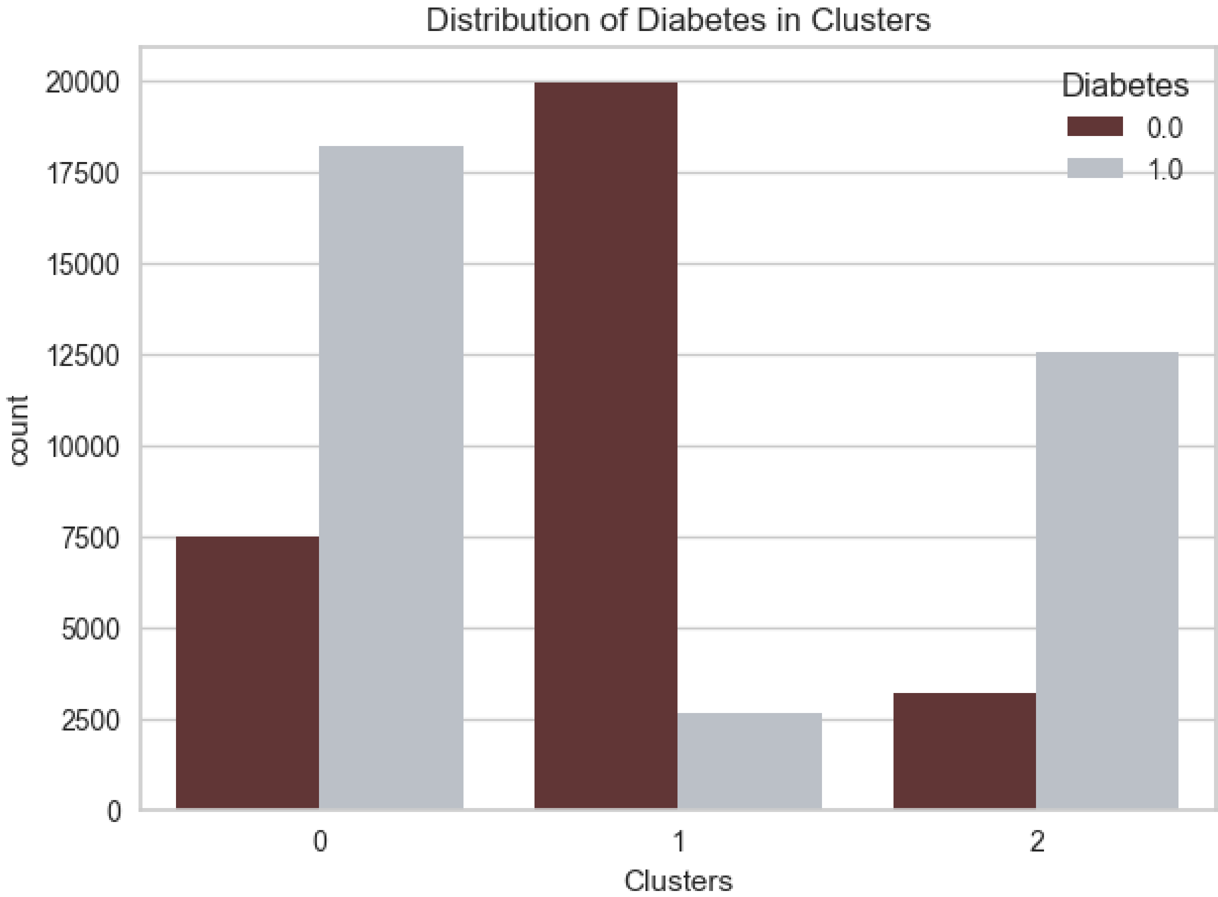



Cluster 0: Most individuals in this cluster have high BP, with approximately 80% reporting high BP and 20% reporting no high BP.

Cluster 1: This cluster has most individuals (approximately 72%) who do not have high BP, with 28% reporting high BP.

Cluster 2: Most individuals in this cluster have high BP, with approximately 80% reporting high BP and 20% reporting no high BP.

### 4.5.5 Diabetes




Cluster 0: Most individuals in this cluster have diabetes, with approximately 69% reporting diabetes and 31% reporting no diabetes.

Cluster 1: This cluster has most individuals (approximately 80%) who do not have diabetes, with 20% reporting diabetes.

Cluster 2: Most individuals in this cluster have diabetes, with approximately 75% reporting diabetes and 25% reporting no diabetes.

5. Overall comparison of clusters

| Category | Cluster 0 | Cluster 1 | Cluster 2 |
|---|---|---|---|
| **Cluster Size** | ~25,000 members | ~22,000 members | ~15,000 members |
| **Age** | Older individuals, median age 65-69 | Middle-aged, median age 45-49 | Older middle-aged, median age 55-59 |
| **BMI** | Mostly "Overweight" to "Obesity II" | "Normal weight" to "Obesity I", median "Overweight" | "Overweight" to "Obesity II" |
| **Gender** | Predominantly male | Predominantly female | More females, but balanced |
| **Physical Activity** | Majority engaged in physical activity | Majority engaged in physical activity | More individuals not engaging in physical activity |
| **Fruits Consumption** | Majority consume fruit daily | Majority consume fruit daily | Balanced daily fruit consumption |
| **Veggies Consumption** | Majority consume vegetables daily | Majority consume vegetables daily | Majority consume vegetables daily, less pronounced |
| **High Alcohol Consumption** | ~10% heavy drinkers | ~20% heavy drinkers | ~5% heavy drinkers |
| **Smoker** | ~53% have smoked | ~39% have smoked | ~56% have smoked |
| **General Health** | Median "Good", IQR "Very Good" to "Fair" | Median "Good", IQR "Very Good" to "Fair" | Median "Fair", IQR "Good" to "Poor" |
| **Mental Health** | Median ~0 days of poor mental health, IQR 0-1 day | Median ~1 day, IQR 0-2 days | Median ~3 days, IQR 1-4 days |
| **Physical Health** | Median ~1 day of physical illness/injury, IQR 0-2 days | Median ~1 day, IQR 0-2 days | Median ~3 days, IQR 1-6 days |

| | | | |
|---|---|---|---|
| **Difficulty Walking** | ~80% no difficulty, ~20% serious difficulty | ~95% no difficulty, ~5% serious difficulty | ~55% serious difficulty, ~45% no difficulty |
| **Heart Disease** | ~20% diagnosed | ~5% diagnosed | ~33% diagnosed |
| **High Cholesterol** | ~68% have high cholesterol | ~24% have high cholesterol | ~65% have high cholesterol |
| **Stroke** | Significant proportion | Lowest proportion | Highest proportion |
| **High Blood Pressure** | ~80% have high BP | ~28% have high BP | ~80% have high BP |
| **Diabetes** | ~69% have diabetes | ~20% have diabetes | ~75% have diabetes |

## 6. Conclusion

Cluster Overview

Cluster 0: Older, Predominantly Male Group

Demographic Characteristics:

This cluster consists of older individuals, with a median age of 65-69, indicating that it predominantly comprises senior citizens.

The gender distribution shows a higher number of males compared to females.

Health Profile:

A significant portion of this cluster is categorized as overweight to obese (Overweight to Obesity II), suggesting a higher risk of weight-related health issues.

High prevalence of chronic conditions: high cholesterol (68%), high blood pressure (80%), and diabetes (69%).

Notable incidence of heart disease (20%) and stroke.

Lifestyle Habits:

Most individuals engage in physical activity and consume fruits and vegetables daily.

Lower proportion of heavy alcohol consumption (~10%).

General and Mental Health:

Most report good general health, with few days of poor mental or physical health.

Approximately 20% have serious difficulty walking or climbing stairs.

Interpretation: This cluster represents older males who generally maintain a healthy lifestyle but have a high prevalence of chronic conditions. Interventions should focus on managing these chronic diseases and encouraging continued physical activity and healthy eating.

Cluster 1: Middle-Aged, Predominantly Female Group

Demographic Characteristics:

This cluster is characterized by middle-aged individuals, with a median age of 45-49.

Predominantly female, indicating a gender skew towards women in this age group.

Health Profile:

BMI ranges from normal weight to overweight, with a median in the overweight range.

Lower prevalence of chronic conditions: high cholesterol (24%), high blood pressure (28%), and diabetes (20%).

Lowest incidence of heart disease (~5%) and stroke.

Lifestyle Habits:

High engagement in physical activity and high daily consumption of fruits and vegetables.

Higher proportion of heavy drinkers compared to Cluster 0 (~20%).

General and Mental Health:

Report good general health, similar to Cluster 0.

Few days of poor mental or physical health.

Over 95% do not have serious difficulty walking or climbing stairs.

Interpretation: This cluster consists of middle-aged women who generally lead a healthy lifestyle and have a lower prevalence of chronic diseases. Preventive measures and health promotion activities should focus on maintaining their health status and addressing the higher incidence of heavy drinking.

Cluster 2: Older Middle-Aged with Balanced Gender Distribution

Demographic Characteristics:

The smallest cluster, with a median age of 55-59, indicating an older middle-aged group.

Balanced gender distribution but with a slight skew towards females.

Health Profile:

High prevalence of overweight and obesity (Overweight to Obesity II).

High proportions of chronic conditions: high cholesterol (65%), high blood pressure (80%), and diabetes (75%).

Higher incidence of heart disease (33%) and the highest proportion with a history of stroke.

Lifestyle Habits:

Less engagement in physical activity compared to the other clusters.

Balanced fruit consumption, with less pronounced daily vegetable consumption.

Lowest heavy alcohol consumption (~5%).

General and Mental Health:

Median health status is "Fair," with a higher number of poor mental and physical health days.

Higher proportion (~55%) with serious difficulty walking or climbing stairs.

Interpretation: This cluster represents an older middle-aged group with a high burden of chronic diseases and lower levels of physical activity. They experience more days of poor mental and physical health and face mobility challenges. Interventions should focus on chronic disease management, mental health support, and promoting physical activity.

Insights

Age and Health Correlation:

Older individuals (Cluster 0) tend to have more chronic health issues and report good health but have a higher prevalence of conditions like diabetes and high blood pressure.

Middle-aged individuals (Cluster 1) exhibit better overall health metrics, suggesting that preventive measures and healthy lifestyle choices are more effective in this age group.

Impact of Lifestyle Choices:

Clusters with higher physical activity and better diet (Clusters 0 and 1) show better health outcomes and fewer days of poor mental and physical health.

Cluster 2, which is less active and has a higher prevalence of health issues, highlights the critical impact of an active lifestyle on overall health.

Gender Distribution and Health:

The gender distribution impacts health outcomes, with Cluster 0 being male-dominated and showing a different health profile compared to female-dominated Cluster 1.

Balanced gender distribution in Cluster 2 suggests that health issues are more related to lifestyle and age rather than gender alone.

Chronic Conditions and Perceived Health:

Despite having a high prevalence of chronic conditions, individuals in Cluster 0 and Cluster 1 perceive their health positively, indicating effective management or a positive health outlook.

Cluster 2's higher self-reported days of poor health and fair health perception align with their higher prevalence of serious health conditions.

Preventive Health Measures:

The lower incidence of heart disease and stroke in Cluster 1 emphasizes the importance of preventive health measures, such as regular exercise, healthy diet, and moderate alcohol consumption.

High levels of diabetes and high blood pressure in Clusters 0 and 2 underscore the need for targeted interventions in these groups to manage and reduce the prevalence of these conditions.

Recommendations

For Cluster 0: Enhance chronic disease management programs focusing on diabetes, high blood pressure, and high cholesterol. Encourage continued physical activity and balanced diet to maintain health.

For Cluster 1: Maintain and promote preventive health measures and healthy lifestyle choices to sustain low levels of chronic conditions and good overall health.

For Cluster 2: Implement targeted interventions to increase physical activity and improve diet. Focus on managing chronic diseases and improving mental and physical health outcomes.

By understanding the distinct characteristics and needs of each cluster, tailored health programs and policies can be developed to improve health outcomes across different demographic groups.