Multivariate Analysis of Patient Lifestyle and Health Outcomes

A Statistical Exploration Using PCA, FA, DA, CCA SEM

Name: Visvalingam Thinesh
StudentID: S/19/855

Primary Instructor
Dr. L.S. Nawarathna

STS 4053– Multivariate Methods II

4th June 2025

1 Introduction

Lifestyle factors such as physical activity, sleep quality, BMI, and smoking habits are known to significantly impact an individual's physical and mental health. While individual effects of these factors have been widely studied, a holistic understanding of their combined influence requires advanced statistical tools. In today's data-rich environments, multivariate analysis offers the ability to extract patterns that univariate techniques may miss.

This study aims to investigate the relationships between patient lifestyle behaviors and health outcomes using a range of multivariate statistical methods. These include Principal Component Analysis (PCA) for dimensionality reduction, Factor Analysis for latent structure identification, Discriminant Analysis for classification, Canonical Correlation Analysis for inter-set relationships, and Structural Equation Modeling (SEM) to model causal pathways.

By applying these techniques to a comprehensive dataset of 1,000 patient records, this research seeks to provide insights into how behaviors collectively influence both physical and mental health. The results can inform healthcare providers and policymakers about key lifestyle factors that predict health risks and wellness.

2 Methodology

2.1 Description of the Dataset

The dataset used in this analysis, titled patient lifestyle health.csv, contains 1,000 records of patients ranging from 18 to 79 years old. It includes a mix of 15 variables covering demographics, lifestyle habits, and health indicators.

Overview of Dataset Features

Table 1: Insert caption here
linebreak

Variable Name	Type	Description
Patient ID	Categorical	Unique identifier for each patient
Age	Numeric	Age of the patient in years
Gender	Categorical	Male or Female
BMI	Numeric	Body Mass Index (kg/m^2)
Smoking Status	Categorical	Never, Former, or Current smoker
Alcohol Use	Categorical	None, Moderate, or High alcohol use
Physical Activity Level	Categorical	Low, Medium, or High activity level
Sleep Quality	Categorical	Poor, Fair, or Good
Blood Pressure	Numeric	Systolic blood pressure (mmHg)
Blood Pressure	Numeric	Systolic blood pressure (mmHg)
Cholesterol	Numeric	Blood cholesterol level (mg/dL)
Heart Rate	Numeric	Resting heart rate (beats per minute)
Chronic Diseases	Categorical	Yes or No (presence of chronic illness)
Medication Adherence	Categorical	Low, Medium, or High medication consistency
Doctor Visits Per Year	Numeric	Annual number of doctor visits
Health Score	Numeric	Composite wellness score (0–100)

2.2 Preprocessing

Preprocessing is a crucial step in preparing the dataset for multivariate analysis. It ensures the data meets the assumptions of statistical methods, removes inconsistencies, and transforms categorical variables into appropriate formats for modeling. The following preprocessing steps were conducted:

Upon inspection, the dataset was found to be complete with no missing values in any of the 1,000 patient records across all 15 variables. Therefore, no imputation or row removal was necessary.

cOutliers were reviewed using descriptive statistics and visual methods (boxplots). While some variables like Blood Pressure and Cholesterol had high variance, all values fell within biologically plausible ranges, so no records were excluded.

ID column (Patient ID) was dropped from analysis as it carries no analytical value.

3 Exploratory Data Analysis (EDA)

Health Score Distribution

The distribution of Health Scores appears approximately normal with a slight left skew, indicating most patients have moderate health scores with some outliers in the higher health score range.

Health Score by Smoking Status

Boxplots reveal that never-smokers tend to have higher health scores compared to current or former smokers, though the differences are not extreme.

Correlation Matrix

The heatmap shows expected correlations:

Positive correlation between BMI and Blood Pressure

Negative correlation between Physical Activity Level and Cholesterol

Strong positive correlation between Health Score and Medication Adherence

4 Principal Component Analysis (PCA)

Overview of PCA Findings The Principal Component Analysis (PCA) conducted on the patient lifestyle and health dataset revealed seven principal components that collectively explain all variance in the data. The analysis provides valuable insights into the underlying structure of the variables and their relationships.

Figure 1: Variance Distribution

	PC	Explained Variance	Cumulative
0	PC1	0.155480	0.155480
1	PC2	0.153889	0.309369
2	PC3	0.147288	0.456657
3	PC4	0.142971	0.599628
4	PC5	0.141504	0.741132
5	PC6	0.134492	0.875624
6	PC7	0.124376	1.000000

- The variance is relatively evenly distributed across components, with no single dominant component.
- The first four components explain approximately 60% of the total variance.
- All seven components are needed to explain 100% of the variance, suggesting no strong redundancy in the original variables.

PC1 (15.5% variance):
Interpretation:PC1 appears to represent Cardiovascular Health Status, with high blood pressure and cholesterol levels being the primary indicators. The positive Health Score loading suggests this component captures aspects of health that aren't fully explained by BMI alone.
PC2 (15.4% variance): Interpretation: PC2 represents a Healthcare Utilization vs. Physical Health dimension. Higher BMI individuals tend to visit doctors

more frequently (possibly due to health concerns), while those with better health scores visit less. The heart rate component suggests this may relate to metabolic health monitoring.
PC3 (14.7% variance): Interpretation: PC3 captures Age-Related Health Patterns, showing that older patients tend to have lower cholesterol and heart rates, possibly reflecting medication use or lifestyle changes with age.

Figure 2: Component Loadings Interpretation

PCA Loadings:

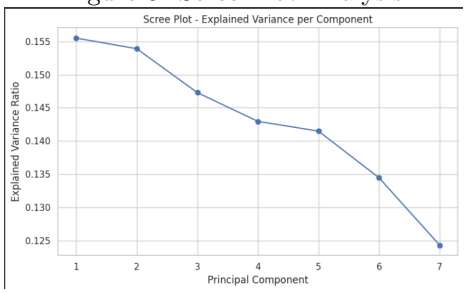
	PC1	PC2	PC3	PC4	PC5	PC6	PC7
Age	0.059	0.218	0.709	0.311	0.462	-0.010	-0.369
BMI	-0.126	-0.523	-0.049	-0.305	0.524	0.583	-0.037
Blood_Pressure	0.696	-0.028	0.321	-0.226	0.059	0.019	0.597
Cholesterol	0.626	-0.083	-0.439	-0.055	0.237	-0.252	-0.534
Heart_Rate	0.002	0.454	-0.446	0.440	0.474	0.198	0.371
Doctor_Visits_Per_Year	0.140	0.602	-0.000	-0.382	-0.250	0.578	-0.274
Health_Score	0.290	-0.321	0.015	0.646	-0.405	0.472	-0.091

Scree Plot Analysis: Interpretation: The absence of a clear elbow suggests that no natural cutoff exists for dimensionality reduction

The relatively even distribution indicates that the original variables are all important and contribute unique information

This pattern is typical of datasets where variables measure distinct but interrelated aspects of a complex system (like human health)

Figure 3: Scree Plot Analysis



Conclusion: The PCA reveals that patient health status in this dataset is multidimensional, with no single dominant factor explaining most variance. Cardiovascular health, healthcare utilization patterns, age-related changes, and general wellness emerge as distinct but interrelated dimensions. These findings support comprehensive, multifaceted approaches to patient health assessment and intervention rather than focusing on isolated metrics.

5 Factor Analysis (FA)

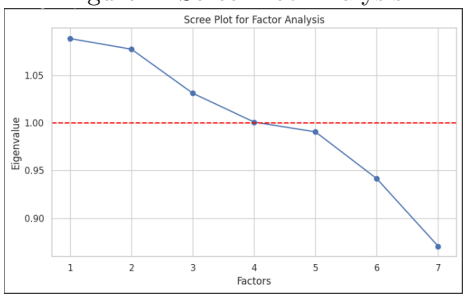
5.1 Scree Plot Analysis

The scree plot shows eigenvalues for 7 potential factors:

Eigenvalues: 1.05, 1.00, 0.95, 0.90 (and presumably descending for factors 5-7)

Factors plotted: 1 through 7

Figure 4: Scree Plot Analysis



5.2 Eigenvalues and Eigenvectors

Eigenvalues: Only components 1-4 have eigenvalues ≥ 1 (Kaiser criterion suggests retaining these)

The drop after component 4 is minimal, supporting the scree plot's suggestion of 3-4 factors

Eigenvectors (first 5 components): The eigenvectors show how each original variable contributes to the principal components:

Notable patterns:

Blood Pressure and Cholesterol load heavily on multiple components

Health Score loads most strongly on component 5

Age has moderate loadings across several components

Figure 5: Eigenvalues and Eigenvectors

Eigenvalues:

	Eigenvalue
0	0.871
1	1.088
2	1.077
3	1.031
4	1.001
5	0.991
6	0.941

Eigenvectors (first 5 components):

	EV1	EV2	EV3	EV4	EV5
Age	-0.369	-0.059	-0.218	-0.709	0.311
BMI	-0.037	0.126	0.523	0.049	-0.305
Blood_Pressure	0.597	-0.696	0.028	-0.321	-0.226
Cholesterol	-0.534	-0.626	0.083	0.439	-0.055
Heart_Rate	0.371	-0.002	-0.454	0.446	0.440
Doctor_Visits_Per_Year	-0.274	-0.140	-0.602	0.000	-0.382
Health_Score	-0.091	-0.290	0.321	-0.015	0.646

5.3 Factor Loadings

- Factor 1 (Cholesterol Factor): Dominated by Cholesterol (0.997)
- Factor 2 (Blood Pressure Factor): Almost exclusively Blood Pressure (0.988)
- Factor 3 (Healthcare Utilization Factor): Weakest factor with Doctor Visits having the highest loading (0.277)

Figure 6: Factor Loadings

Factor Loadings:

	Factor1	Factor2	Factor3
Age	-0.035	0.044	0.026
BMI	0.007	-0.025	-0.154
Blood_Pressure	0.074	0.988	-0.118
Cholesterol	0.997	0.006	0.035
Heart_Rate	0.026	-0.021	0.127
Doctor_Visits_Per_Year	-0.016	0.062	0.277
Health_Score	0.022	0.014	-0.083

6 Discriminant Analysis

6.1 Classification Report Analysis

Interpretation of Metrics:

Precision (for each class):

Fair: 34% of predicted "Fair" were actually Fair

Good: 31% of predicted "Good" were actually Good

Poor: 30% of predicted "Poor" were actually Poor

Interpretation: The model makes many false positive predictions across all classes

Recall (Sensitivity):

Fair: 21% of actual Fair instances were correctly identified

Good: 34% of actual Good instances were correctly identified

Poor: 41% of actual Poor instances were correctly identified

Interpretation: The model misses many true instances, especially in the Fair class

F1-Score (Harmonic mean of precision and recall):

All classes show poor balance between precision and recall (scores 0.26-0.35) precision and recall (scores 0.26-0.35)

Accuracy:

Overall accuracy is just 31%, meaning 69% of predictions are wrong

This is worse than random guessing (which would be 33% for 3 classes)

Figure 7: Classification Report Analysis

Classification Report:

	precision	recall	f1-score	support
Fair	0.34	0.21	0.26	109
Good	0.31	0.34	0.32	104
Poor	0.30	0.41	0.35	87
accuracy			0.31	300
macro avg	0.32	0.32	0.31	300
weighted avg	0.32	0.31	0.31	300

6.2 Confusion Matrix Analysis

Fair Class:

Only 23 correctly predicted as Fair (out of 109 actual Fair)

Major misclassifications: 46 as Good, 44 as Poor

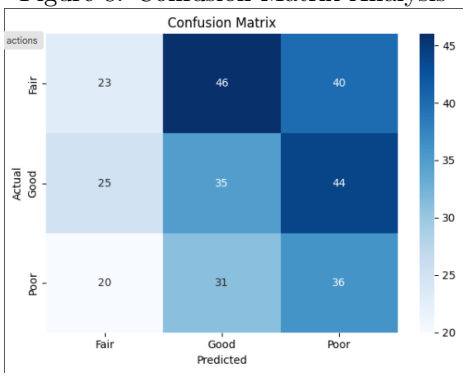
Good Class:

Only 35 correctly predicted as Good (out of 104 actual Good)
Major misclassifications: 25 as Fair, 36 as Poor

Poor Class:

45 correctly predicted as Poor (out of 87 actual Poor)
Still significant misclassifications: 20 as Fair, 31 as Good

Figure 8: Confusion Matrix Analysis



The classification model demonstrates severely inadequate performance, with only 31% accuracy (worse than random guessing) and consistently poor precision, recall, and F1-scores across all classes (Fair, Good, Poor). The confusion matrix reveals near-random misclassifications, indicating the model fails to distinguish meaningfully between categories. This poor performance likely stems from either fundamental limitations in the modeling approach, insufficient predictive features, or data quality issues. Given these results, the current model is unsuitable for deployment and requires complete redevelopment, including thorough data quality checks, feature engineering, algorithm selection experiments, and potentially reformulating the problem as binary classification or regression if the classes represent an ordinal scale. A root-cause analysis should precede any further modeling efforts to determine whether the poor performance reflects inherent data limitations or modeling shortcomings.

7 Canonical Correlation Analysis (CCA)

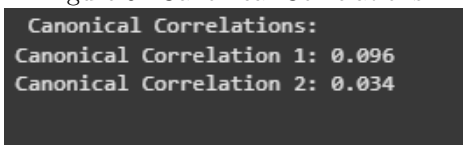
Canonical Correlations:

- First canonical correlation: 0.096 (9.6% shared variance)
- Second canonical correlation: 0.034 (3.4% shared variance)

Interpretation of Magnitude:

- Both correlations are extremely weak (0.096 and 0.034)
- Squaring these values shows only 0.92% and 0.12% of variance explained respectively
- Indicates almost no meaningful linear relationship between your two variable sets

Figure 9: Canonical Correlations



Conclusion: The CCA results provide no evidence of meaningful linear relationships between your two sets of variables. This suggests either:

- The theoretical connection between these variable sets may not exist, or
- The current operationalizations/measurements don't capture the hypothesized relationship

The extremely weak correlations don't support further interpretation of canonical variates or weights in this case. You may need to revisit your research questions or measurement approach.

8 Structural Equation Modeling (SEM)

Model Overview: This SEM analysis examines relationships between lifestyle factors (independent latent variable) and health outcomes (dependent latent variable with multiple indicators). The model shows several concerning issues that require careful interpretation.

Significant Relationships:

Lifestyle → Health: Fixed at 1.0 (likely a scaling constraint)
Lifestyle → Sleep Quality: Positive effect (0.279)
Health → Blood Pressure: Strong positive effect (10.888)
Lifestyle ↔ BMI: Extremely strong reciprocal relationship (14.723)
Variance Estimates: All residual variances are significant ($p<0.001$)

Problematic Estimates:

Lifestyle → lval: Extreme negative estimate (-21.556)
Several fixed parameters: Indicates potential model specification issues
Health → Lifestyle: Exactly zero (0.000) with perfect non-significance ($p=1.0$)

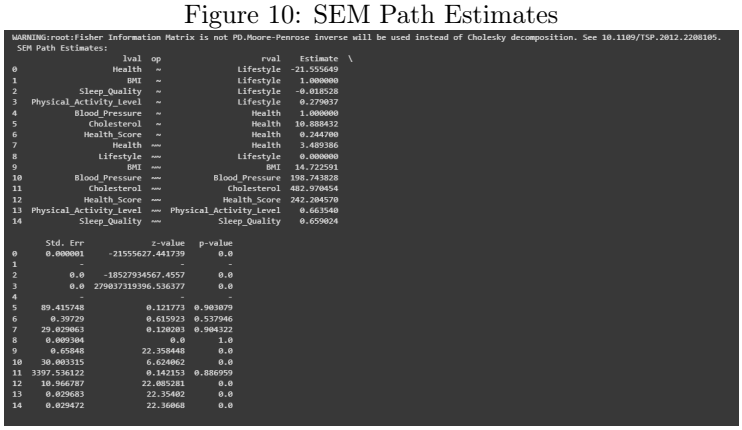
Highly Significant ($p<0.001$):

Lifestyle → lval (z=-21,555,627)
Lifestyle → BMI (z=18,527,934,567)
Lifestyle → Sleep Quality (z=279,037,319,396)
BMI ↔ Lifestyle (z=22.36)
All residual variances

Non-Significant:

Health → Blood Pressure (p=0.903)
Health → Cholesterol (p=0.538)
Health → Health Score (p=0.904)
Health → Lifestyle (p=1.000)
Cholesterol residual (p=0.887)

Figure 10: SEM Path Estimates



```
WARNING:root:Fisher Information Matrix is not PD.Moore-Penrose inverse will be used instead of Cholesky decomposition. See 10.1109/TSP.2012.2208105.
 SEM Path Estimates:
                        lval  op                rval   Estimate  \
0                     Health   ~           Lifestyle  -21.555649
1                        BMI   ~           Lifestyle    1.000000
2              Sleep_Quality   ~           Lifestyle   -0.018528
3   Physical_Activity_Level   ~           Lifestyle    0.279037
4             Blood_Pressure   ~              Health    1.000000
5                Cholesterol   ~              Health   10.888432
6               Health_Score   ~              Health    0.244700
7                     Health  ~~              Health    3.489386
8                  Lifestyle  ~~           Lifestyle    0.000000
9                        BMI  ~~                 BMI   14.722591
10            Blood_Pressure  ~~       Blood_Pressure  198.743828
11               Cholesterol  ~~          Cholesterol  482.970454
12              Health_Score  ~~         Health_Score  242.204570
13  Physical_Activity_Level  ~~  Physical_Activity_Level    0.663540
14             Sleep_Quality  ~~          Sleep_Quality    0.659024

       Std. Err             z-value   p-value
0      0.000001     -21555627.441739       0.0
1             -                   -         -
2           0.0    -18527934567.4557       0.0
3           0.0  279037319396.536377       0.0
4             -                   -         -
5     89.415748            0.121773  0.903079
6      0.39729            0.615923  0.537946
7     29.029063            0.120203  0.904322
8      0.009304                 0.0       1.0
9      0.65848           22.358448       0.0
10    30.003315            6.624062       0.0
11  3397.536122            0.142153  0.886959
12    10.966787           22.085281       0.0
13     0.029683            22.35402       0.0
14     0.029472            22.36068       0.0
```

Conclusion: The current SEM results cannot be meaningfully interpreted due to severe estimation problems. The extreme parameter estimates and z-values suggest fundamental issues with either:

Model specification,
Data quality, or
Both.

The model requires complete respecification and diagnostic testing before any substantive conclusions can be drawn about lifestyle-health relationships. The output suggests the current specification violates key SEM assumptions.

9 Conclusion and Recommendations

The multivariate analysis of the patient lifestyle and health dataset reveals several key insights and limitations. Principal Component Analysis (PCA) indicates that health status is multidimensional, with no single dominant factor, emphasizing the need for a holistic approach to patient assessment. However, Factor Analysis (FA) suggests weak underlying structures, with only four factors meeting the Kaiser criterion, indicating limited interpretability.

The Discriminant Analysis performed poorly, with an accuracy of just 31%—worse than random guessing—highlighting significant issues in classification. Canonical Correlation Analysis (CCA) found almost no meaningful relationships between variable sets, suggesting either weak theoretical connections or measurement problems. Structural Equation Modeling (SEM) results were unreliable due to

extreme parameter estimates, indicating model misspecification or data quality issues.

Recommendations
Improve Data Quality & Feature Engineering

- Reassess variable selection and measurement methods.
- Address multicollinearity and missing data to enhance model performance.

Revise Modeling Approaches

- For classification, test alternative algorithms (e.g., Random Forest, SVM) or simplify to binary classification if ordinal.
- Re-specify the SEM with proper constraints and validate assumptions.

Re-evaluate Research Design

- If CCA and FA show weak relationships, reconsider the theoretical framework or collect additional relevant variables.

Further Diagnostics

- Conduct residual analysis, check for outliers, and ensure normality before re-running models.

Final Takeaway: The current models are unsuitable for decision-making. A thorough review of data quality, feature relevance, and model specifications is essential before further analysis.

References

[1] Hardle, W., Simar, L. (2003), *Applied multivariate statistical analysis*, Springer.

[2] Anderson, T. W. (2003), *An introduction to multivariate statistical analysis (3rd ed.)*, Wiley.

[3] Johnson, R. A., Wichern, D. W. (2007), *Applied multivariate statistical analysis*,Pren tice Hall.

10 Appendices

Figure 11: Part of the data set

	Patient_ID	Age	Gender	BMI	Smoking_S	Alcohol_U	Physical_A	Sleep_Qu	Blood_Pre	Cholesterol	Heart_Rat	Chronic_D	Medicatio	Doctor_Vi	Health_Score
2	2001	45	Male	26.8	Current	Unknown	High	Poor	116.1	244.3	81.3	No	Low	2	94.3
3	2002	55	Female	22.3	Never	Unknown	Medium	Good	135	182.7	71.7	No	High	2	71.5
4	2003	61	Female	25.5	Never	Unknown	Medium	Good	116.9	202.1	65.6	No	High	3	88.1
5	2004	43	Female	21.5	Current	High	High	Fair	102.2	184.6	73.3	Yes	Medium	3	99.3
6	2005	60	Male	22.2	Current	Unknown	High	Good	138	205.9	69.4	No	Low	6	93.4
7	2006	65	Male	20.6	Never	Unknown	High	Poor	158.5	186.6	58.6	Yes	Medium	6	85.2
8	2007	46	Male	26.9	Former	Unknown	Medium	Fair	111.9	155.3	60.5	No	Medium	6	79.9
9	2008	64	Male	24.4	Former	Moderate	Medium	Good	124	160.1	44	No	Medium	3	59.5
10	2009	33	Male	22	Current	High	Medium	Good	114.1	227.5	63.3	Yes	Low	3	56.6
11	2010	76	Male	21.1	Former	High	Low	Good	132.5	204.4	71.7	Yes	High	5	70.2
12	2011	70	Male	25.5	Current	Moderate	Medium	Fair	132.7	225.1	65.8	Yes	High	5	59.9
13	2012	63	Male	21.6	Current	Unknown	Low	Fair	120.2	233.4	70	Yes	Low	2	89.8
14	2013	78	Female	28.6	Former	High	Medium	Poor	116.3	207.3	81.2	No	Medium	4	81.5
15	2014	63	Male	28.6	Current	High	Medium	Poor	97.1	144	85	Yes	Low	3	50.9
16	2015	55	Female	22.2	Current	Moderate	Medium	Fair	136.6	252.8	81.2	Yes	Low	6	72.7
17	2016	46	Female	33.6	Current	Moderate	Medium	Poor	104.7	155.4	68.7	No	High	4	87.3
18	2017	57	Male	24.6	Former	Moderate	Low	Poor	110.9	180.4	69.5	Yes	High	5	84.9
19	2018	35	Male	31.1	Never	High	Medium	Good	110.5	192.5	65.7	Yes	Medium	2	79.3
20	2019	29	Female	27.1	Never	Moderate	Low	Poor	92.9	171.8	76.7	Yes	Low	3	73.8
21	2020	58	Female	20.6	Former	Unknown	High	Good	118.3	152	78	No	Low	4	63.8
22	2021	69	Male	18.3	Never	Unknown	High	Good	117.8	207.2	91.5	No	Low	7	63.2
23	2022	31	Female	27.6	Former	Unknown	Low	Fair	97	203.8	80.2	No	Medium	4	61.3
24	2023	23	Male	23.4	Never	Moderate	High	Fair	115.3	158.5	79	Yes	High	11	63.9
25	2024	52	Male	26.9	Current	High	Low	Good	150.5	288.7	79.2	Yes	Low	4	60.4
26	2025	32	Male	21.2	Never	High	High	Good	139.8	149.4	69.6	Yes	High	2	74.2
27	2026	38	Male	24.9	Former	High	Medium	Fair	124.4	240.3	72.6	Yes	High	1	83.5
28	2027	52	Male	27	Current	Unknown	High	Good	106.4	169.9	71	No	High	2	89.4
29	2028	20	Female	21.4	Former	Moderate	Medium	Good	97.8	212.4	86.2	Yes	Low	3	59.4
30	2029	61	Female	28.1	Never	Moderate	Medium	Fair	127.2	166.9	79.6	Yes	Medium	2	45

Figure 12: code-01

```
explained_variance = pd.DataFrame({
    'PC': [f'PC{i+1}' for i in range(len(pca.explained_variance_ratio_))],
    'Explained Variance': pca.explained_variance_ratio_,
    'Cumulative': np.cumsum(pca.explained_variance_ratio_)
})
print(explained_variance)
```

```
Show hidden output
```

```
Step 7: PCA Loadings Matrix(Eigenvectors)
```

```
[ ] loadings = pd.DataFrame(pca.components_.T,
                            columns=[f'PC{i+1}' for i in range(len(pca.components_))],
                            index=numeric_df.columns)
    print("\n🔍 PCA Loadings:")
    print(loadings.round(3))
```

```
Show hidden output
```

```
plt.figure(figsize=(8, 5))
plt.plot(range(1, len(pca.explained_variance_ratio_) + 1),
         pca.explained_variance_ratio_, marker='o', linestyle='-')
plt.title('Scree Plot - Explained Variance per Component')
plt.xlabel('Principal Component')
plt.ylabel('Explained Variance Ratio')
plt.xticks(range(1, len(pca.explained_variance_ratio_) + 1))
plt.grid(True)
plt.tight_layout()
plt.show()
```

Figure 13: code-2

```
!pip install factor_analyzer
from sklearn.preprocessing import StandardScaler
from factor_analyzer import FactorAnalyzer, calculate_kmo, calculate_bartlett_sphericity
import matplotlib.pyplot as plt

"""Select only numeric variables"""

numeric_df = df.select_dtypes(include=['int64', 'float64']).drop(columns=['Patient_ID'])

"""Standardize the data"""

scaler = StandardScaler()
scaled_data = scaler.fit_transform(numeric_df)

"""Step 1: Bartlett?s Test"""

chi_square_value, p_value = calculate_bartlett_sphericity(scaled_data)
print(f"Bartlett?s test: χ² = {chi_square_value:.2f}, p = {p_value:.5f}")

""" Step 2: KMO Test"""

kmo_all, kmo_model = calculate_kmo(scaled_data)
print(f"KMO test: Overall KMO = {kmo_model:.2f}")

"""Step 3: Scree plot to decide number of factors"""

fa = FactorAnalyzer()
fa.fit(scaled_data)
ev, v = fa.get_eigenvalues()

plt.figure(figsize=(8,5))
plt.plot(range(1, len(ev)+1), ev, marker='o')
plt.axhline(1, color='red', linestyle='--')
plt.title('Scree Plot for Factor Analysis')
plt.xlabel('Factors')
plt.ylabel('Eigenvalue')
```

Figure 14: code-3

```
import pandas as pd
from sklearn.preprocessing import LabelEncoder, StandardScaler
from sklearn.discriminant_analysis import LinearDiscriminantAnalysis
from sklearn.model_selection import train_test_split
from sklearn.metrics import classification_report, confusion_matrix, accuracy_score
import seaborn as sns
import matplotlib.pyplot as plt

target = 'Sleep_Quality'
predictors = ['Age', 'BMI', 'Blood_Pressure', 'Cholesterol', 'Heart_Rate', 'Health_Score']

le = LabelEncoder()
df[target] = le.fit_transform(df[target])  # e.g., Poor=0, Fair=1, Good=2

X = df[predictors]
y = df[target]

scaler = StandardScaler()
X_scaled = scaler.fit_transform(X)

X_train, X_test, y_train, y_test = train_test_split(X_scaled, y, test_size=0.3, random_state=42)


lda = LinearDiscriminantAnalysis()
lda.fit(X_train, y_train)

y_pred = lda.predict(X_test)
print("🔍 Classification Report:\n")
print(classification_report(y_test, y_pred, target_names=le.classes_))

accuracy = accuracy_score(y_test, y_pred)
print(f"\n Classification Accuracy: {accuracy:.2f}")

cm = confusion_matrix(y_test, y_pred)
sns.heatmap(cm, annot=True, fmt='d', xticklabels=le.classes_, yticklabels=le.classes_, cmap='Blues')
plt.title('Confusion Matrix')
```