Multivariate Analysis of Student Habits and Academic Performance

Course: STA4053 - Multivariate Methods II
Name: D.R.K.Dhanushka
Reg. Number: S/19/809

1. Introduction

This study investigates how students' daily habits and lifestyle choices affect their academic performance. With the increasing influence of technology, mental health, and socio-environmental factors, identifying key patterns among students is essential. The primary objective is to use multivariate statistical techniques to uncover latent structures and relationships within a comprehensive dataset of 1000 students, ultimately guiding educational strategies and interventions.

2. Methodology

2.1 Dataset Description

The dataset comprises 1000 student records, each with 16 variables:

- **Continuous Variables:** Age, Study Hours per Day, Social Media Hours, Netflix Hours, Attendance Percentage, Sleep Hours, Exercise Frequency, Mental Health Rating, Exam Score
- **Categorical Variables:** Gender, Part-time Job, Diet Quality, Parental Education Level, Internet Quality, Extracurricular Participation

2.2 Preprocessing Steps

- Removed non-analytical ID field (student_id)
- Dropped records with missing values to ensure complete cases for analysis

- Plotted histograms of all numerical variables to assess distribution shapes and detect skewness
- Applied log transformations to variables with high skewness to normalize distributions
- Plotted boxplots of numerical features to identify potential outliers
- Removed outliers using the IQR method to clean the data
- Plotted a correlation heatmap to identify linear relationships and multicollinearity among continuous variables
- Converted categorical variables to dummy/indicator variables for analysis
- Standardized continuous variables for PCA and clustering

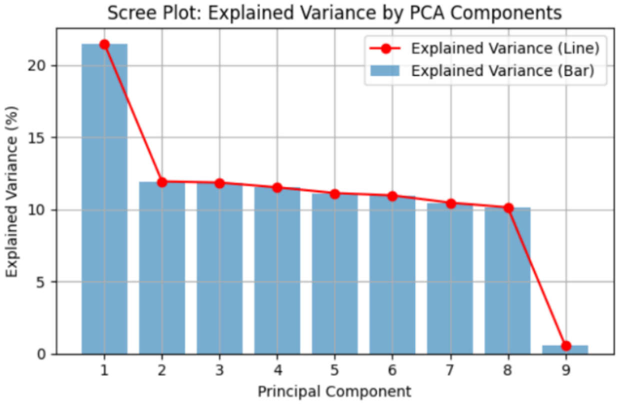
Note : Although log transformations were applied to reduce skewness in highly skewed variables, some remained moderately skewed. Given the robustness of multivariate techniques to such deviations, analysis proceeded with standardized data.

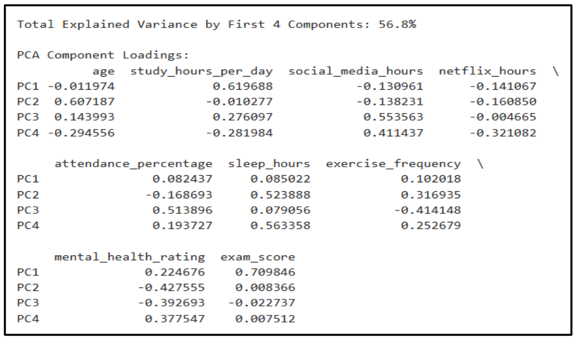
2.3 Multivariate Techniques Applied

- **Principal Component Analysis (PCA):** To reduce dimensionality and identify key underlying components.
- **Cluster Analysis (K-Means):** To segment students into behavioral or performance-based groups.
- **Discriminant Analysis (LDA):** To predict exam performance based on lifestyle indicators.
- **MANOVA:** To analyze how categorical factors influence academic and mental health outcomes.

3. Results and Discussion

3.1 Principal Component Analysis (PCA)





```
Total Explained Variance by First 4 Components: 56.8%

PCA Component Loadings:
          age  study_hours_per_day  social_media_hours  netflix_hours  \
PC1 -0.011974             0.619688           -0.130961      -0.141067
PC2  0.607187            -0.010277           -0.138231      -0.160850
PC3  0.143993             0.276097            0.553563      -0.004665
PC4 -0.294556            -0.281984            0.411437      -0.321082

     attendance_percentage  sleep_hours  exercise_frequency  \
PC1               0.082437     0.085022            0.102018
PC2              -0.168693     0.523888            0.316935
PC3               0.513896     0.079056           -0.414148
PC4               0.193727     0.563358            0.252679

     mental_health_rating  exam_score
PC1              0.224676    0.709846
PC2             -0.427555    0.008366
PC3             -0.392693   -0.022737
PC4              0.377547    0.007512
```

- PCA, with the first four components, explains approximately 56.8% of the total variance.

- While this captures a meaningful portion of the data's variability, it falls short of the typical 70–80% threshold expected for dimensionality reduction in multivariate analysis.

- As such, the interpretation of the components should be approached with caution, and additional components may be explored in future analyses if higher cumulative variance is needed.

Interpretation of Principal Components: The names assigned to each component are based on the strongest variable loadings observed in the PCA loadings matrix. These names are descriptive summaries intended to reflect the shared thematic meaning of the variables contributing most to each component.

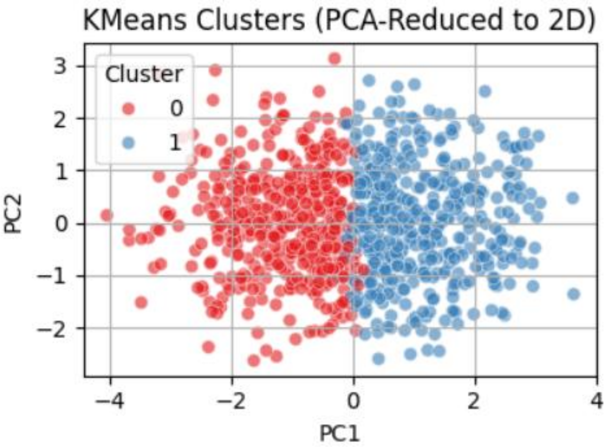
- **PC1 (Academic Commitment and Performance):** High positive loadings on study_hours_per_day and exam_score suggest that this component represents students who dedicate time to studying and perform well academically.

- **PC2 (Lifestyle and Physical Wellness):** Strong contributions from sleep_hours and exercise_frequency, and a negative relationship with mental_health_rating, reflect a lifestyle dimension where rest and activity levels are central.

- **PC3 (Engagement and Digital Distraction):** Driven by high social_media_hours and positive attendance_percentage, this component may reflect the tension between classroom involvement and digital distractions.
- **PC4 (Routine Balance and Wellness):** Moderate loadings from sleep_hours, mental_health_rating, and social_media_hours indicate a subtler axis that reflects students' overall routine balance and psychological well-being.

Despite the moderate explained variance, these components still help simplify complex student behaviors into understandable dimensions that can support clustering and classification tasks. That can be further used in clustering and predictive modeling.

3.2 Cluster Analysis



- To identify natural groupings among students based on their standardized behavior and performance data, K-Means clustering was applied.
- The optimal number of clusters was determined to be **2**, based on the silhouette score which measures how well each data point fits within its assigned cluster.
- Visualization of the clusters was achieved by reducing the data to two principal components using PCA, making it easier to interpret the distribution and separation of clusters.

Cluster Interpretations: To meaningfully understand each cluster, we computed the average of important behavioral and academic variables per group. Because variables were standardized, the values are in terms of standard deviations (SD) from the mean. The summary statistics helped characterize the clusters based on actual student attributes.

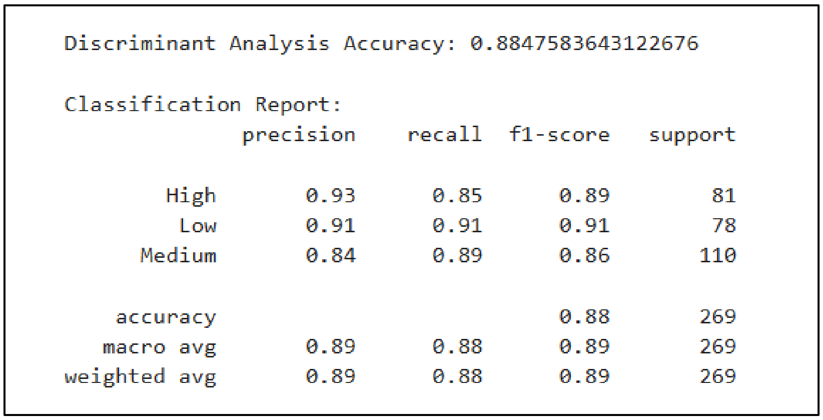
Cluster	Study Hours	Social media	Netflix	Attendance	Sleep	Mental Health	Exam Score
0	-0.65	0.17	0.17	-0.09	-0.17	-0.33	-0.81
1	0.64	-0.16	-0.16	0.09	0.17	0.32	0.79

- **Cluster 0:** These students tend to study less, spend more time on social media and Netflix, sleep less, and report lower mental health and academic scores. They represent the **lower-performing group**.
- **Cluster 1:** This group demonstrates stronger academic behaviors, with more study time, better sleep and mental health, and higher exam scores. They represent the **higher-performing students**.

This analysis highlights behavioral patterns that align with academic outcomes and can inform targeted interventions or student support strategies. Low performers with high social media/Netflix time and poor sleep.

3.3 Discriminant Analysis

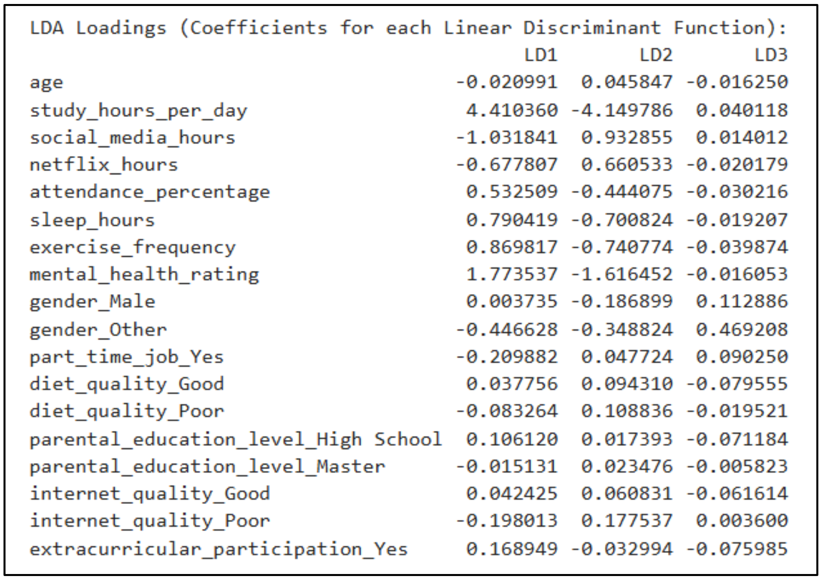
Linear Discriminant Analysis (LDA) was used to predict academic performance groups (Low, Medium, High) based on students' lifestyle and background features.



```
Discriminant Analysis Accuracy: 0.8847583643122676

Classification Report:
              precision    recall  f1-score   support

        High       0.93      0.85      0.89        81
         Low       0.91      0.91      0.91        78
      Medium       0.84      0.89      0.86       110

    accuracy                           0.88       269
   macro avg       0.89      0.88      0.89       269
weighted avg       0.89      0.88      0.89       269
```

- Categorized exam_score into High (>=80), Medium (60–79), Low (<60)
- Accuracy: 88.5% on the test set
- The model shows strong predictive performance, with all three categories of performance (High, Medium, Low) being classified with good balance between precision and recall.
- Classification Performance:
    - **High Performers:** Precision = 0.93, Recall = 0.85, F1-score = 0.89 — indicating most high scorers were correctly identified
    - **Low Performers:** Precision = 0.91, Recall = 0.91, F1-score = 0.91 — the model effectively captured struggling students
    - **Medium Performers:** Precision = 0.84, Recall = 0.89, F1-score = 0.86 — slightly lower precision but strong recall for middle-range students

Key Predictors Identified from LDA Loadings:



```
LDA Loadings (Coefficients for each Linear Discriminant Function):
                                           LD1       LD2       LD3
age                                  -0.020991  0.045847 -0.016250
study_hours_per_day                   4.410360 -4.149786  0.040118
social_media_hours                   -1.031841  0.932855  0.014012
netflix_hours                        -0.677807  0.660533 -0.020179
attendance_percentage                 0.532509 -0.444075 -0.030216
sleep_hours                           0.790419 -0.700824 -0.019207
exercise_frequency                    0.869817 -0.740774 -0.039874
mental_health_rating                  1.773537 -1.616452 -0.016053
gender_Male                           0.003735 -0.186899  0.112886
gender_Other                         -0.446628 -0.348824  0.469208
part_time_job_Yes                    -0.209882  0.047724  0.090250
diet_quality_Good                     0.037756  0.094310 -0.079555
diet_quality_Poor                    -0.083264  0.108836 -0.019521
parental_education_level_High School  0.106120  0.017393 -0.071184
parental_education_level_Master      -0.015131  0.023476 -0.005823
internet_quality_Good                 0.042425  0.060831 -0.061614
internet_quality_Poor                -0.198013  0.177537  0.003600
extracurricular_participation_Yes     0.168949 -0.032994 -0.075985
```

- The most influential predictors in separating the groups were:
    - study_hours_per_day (strongly positive in LD1)
    - mental_health_rating and exercise_frequency
    - sleep_hours, social_media_hours, and attendance_percentage

- These were determined based on their high absolute coefficient values in the first two discriminant functions (LD1 and LD2).

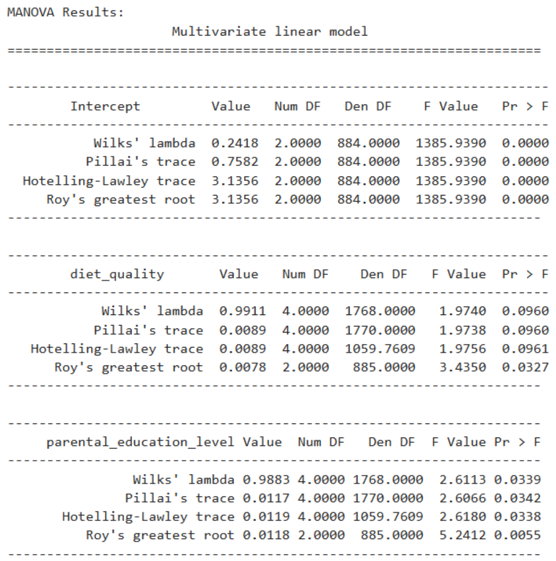
These findings align with expected patterns: Academically successful students tend to spend more time studying, maintain healthier routines, and experience better mental health.

Interpretation: The model's accuracy of 88.5% demonstrated its excellent ability to distinguish between performance levels. LD1 captured separation between high and low performers based largely on study time, mental health, and wellness factors, while LD2 further refined distinctions using similar lifestyle indicators.

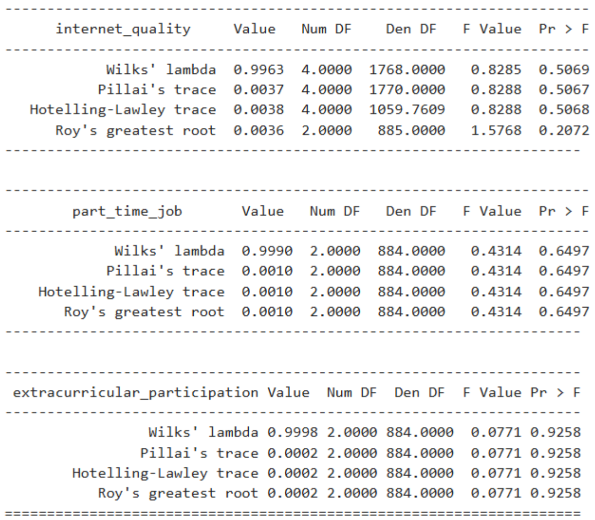
These findings confirm that consistent study habits, mental well-being, and balanced lifestyle factors are reliable student performance indicators and can help guide early academic support.

3.4 MANOVA

MANOVA (Multivariate Analysis of Variance) was used to assess how several categorical variables influence both exam score and mental health rating.



```
MANOVA Results:
                    Multivariate linear model
=================================================================

------------------------------------------------------------------
        Intercept         Value  Num DF  Den DF    F Value  Pr > F
------------------------------------------------------------------
           Wilks' lambda  0.2418 2.0000 884.0000 1385.9390 0.0000
          Pillai's trace  0.7582 2.0000 884.0000 1385.9390 0.0000
  Hotelling-Lawley trace  3.1356 2.0000 884.0000 1385.9390 0.0000
     Roy's greatest root  3.1356 2.0000 884.0000 1385.9390 0.0000
-----------------------------------------------------------------

------------------------------------------------------------------
        diet_quality       Value  Num DF   Den DF   F Value  Pr > F
------------------------------------------------------------------
            Wilks' lambda  0.9911 4.0000 1768.0000  1.9740 0.0960
           Pillai's trace  0.0089 4.0000 1770.0000  1.9738 0.0960
   Hotelling-Lawley trace  0.0089 4.0000 1059.7609  1.9756 0.0961
      Roy's greatest root  0.0078 2.0000  885.0000  3.4350 0.0327
-----------------------------------------------------------------

-----------------------------------------------------------------
     parental_education_level Value  Num DF   Den DF  F Value Pr > F
-----------------------------------------------------------------
               Wilks' lambda 0.9883 4.0000 1768.0000  2.6113 0.0339
              Pillai's trace 0.0117 4.0000 1770.0000  2.6066 0.0342
      Hotelling-Lawley trace 0.0119 4.0000 1059.7609  2.6180 0.0338
         Roy's greatest root 0.0118 2.0000  885.0000  5.2412 0.0055
-----------------------------------------------------------------
```



```
--------------------------------------------------------------------
      internet_quality     Value  Num DF    Den DF   F Value  Pr > F
--------------------------------------------------------------------
            Wilks' lambda  0.9963 4.0000 1768.0000   0.8285 0.5069
           Pillai's trace  0.0037 4.0000 1770.0000   0.8288 0.5067
   Hotelling-Lawley trace  0.0038 4.0000 1059.7609   0.8288 0.5068
      Roy's greatest root  0.0036 2.0000  885.0000   1.5768 0.2072
-------------------------------------------------------------------

--------------------------------------------------------------------
        part_time_job       Value  Num DF   Den DF   F Value  Pr > F
--------------------------------------------------------------------
             Wilks' lambda  0.9990 2.0000 884.0000   0.4314 0.6497
            Pillai's trace  0.0010 2.0000 884.0000   0.4314 0.6497
    Hotelling-Lawley trace  0.0010 2.0000 884.0000   0.4314 0.6497
       Roy's greatest root  0.0010 2.0000 884.0000   0.4314 0.6497
-------------------------------------------------------------------

-------------------------------------------------------------------
 extracurricular_participation Value  Num DF  Den DF  F Value Pr > F
-------------------------------------------------------------------
                 Wilks' lambda 0.9998 2.0000 884.0000  0.0771 0.9258
                Pillai's trace 0.0002 2.0000 884.0000  0.0771 0.9258
        Hotelling-Lawley trace 0.0002 2.0000 884.0000  0.0771 0.9258
           Roy's greatest root 0.0002 2.0000 884.0000  0.0771 0.9258
===================================================================
```

- **Significant Predictor:**
  - Parental Education Level showed a statistically significant multivariate effect (Wilks' λ = 0.9883, F = 2.61, p = 0.034), suggesting students with higher parental education levels tend to perform better academically and report better mental health.
- **Marginal Predictor:**
  - Diet Quality showed a marginal multivariate effect (Wilks' λ = 0.9911, F = 1.97, p = 0.096). Although not statistically significant at the 0.05 level, it may still play a minor role in student outcomes and could be explored further.
- **Non-significant Predictors:**
  - Internet Quality, Part-Time Job, and Extracurricular Participation did not show statistically significant multivariate effects (p > 0.05).

These results highlight the role of family educational background as a contributing factor to student success, while other lifestyle or contextual variables appear to have limited multivariate influence in this dataset.

4. Conclusion and Recommendation

This study demonstrates how multivariate techniques can uncover relationships between student behaviors and academic performance. PCA identified meaningful, though moderately explanatory, behavioral dimensions. Clustering highlighted distinct performance-based student groups, and LDA achieved high accuracy in predicting performance levels. MANOVA revealed that among the contextual factors analyzed, only parental education level had a statistically significant effect on both academic and mental health outcomes.

Key Takeaways:

- Academic engagement (study hours, attendance) and personal wellness (sleep, mental health) are major differentiators between high and low performers.
- PCA helped reduce dimensionality, though it captured slightly less variance than expected, suggesting potential for additional feature inclusion or alternative methods.
- Discriminant Analysis provided robust performance for academic group classification.
- MANOVA confirmed the relevance of parental education level as a contextual factor, while internet quality, part-time jobs, and extracurriculars had limited statistical influence.

Limitations:

- Self-reported data may include bias or inaccuracies.
- Temporal changes (e.g., exam season effects) were not accounted for.

Recommendations:

- Encourage structured study routines and consistent attendance.
- Integrate mental health support and sleep hygiene education into academic programs.
- Use cluster insights to design targeted interventions for struggling students.
- Consider additional behavioral or psychological factors in future data collection to improve model variance explanation.

5. References

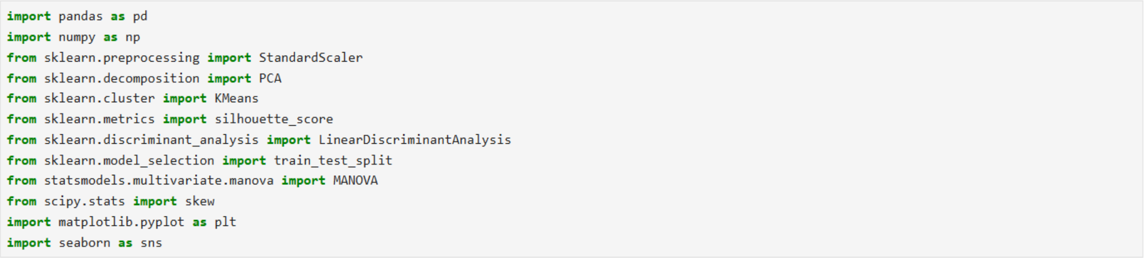
- Hair, J. F., Black, W. C., Babin, B. J., & Anderson, R. E. (2019). *Multivariate Data Analysis* (8th ed.). Cengage Learning.
- Tabachnick, B. G., & Fidell, L. S. (2013). *Using Multivariate Statistics* (6th ed.). Pearson.

6. Appendices

- **Dataset :** https://www.kaggle.com/datasets/jayaantanaath/student-habits-vs-academic-performance

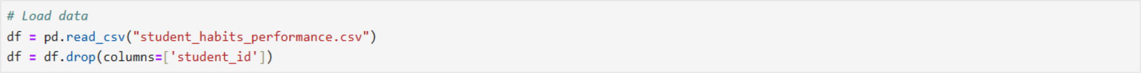
- **Python Code :**

Import Important Libraries

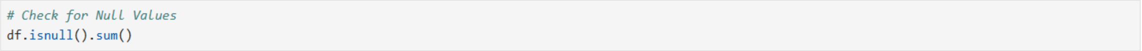


```python
import pandas as pd
import numpy as np
from sklearn.preprocessing import StandardScaler
from sklearn.decomposition import PCA
from sklearn.cluster import KMeans
from sklearn.metrics import silhouette_score
from sklearn.discriminant_analysis import LinearDiscriminantAnalysis
from sklearn.model_selection import train_test_split
from statsmodels.multivariate.manova import MANOVA
from scipy.stats import skew
import matplotlib.pyplot as plt
import seaborn as sns
```

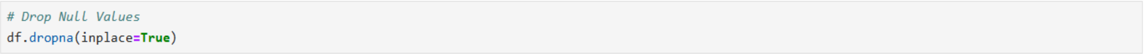
Exploratory Data Analysis



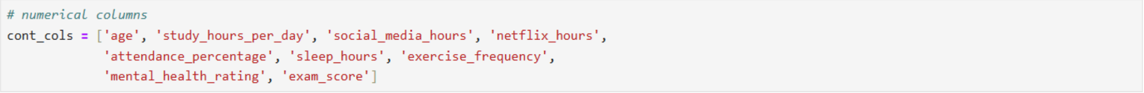
```python
# Load data
df = pd.read_csv("student_habits_performance.csv")
df = df.drop(columns=['student_id'])
```



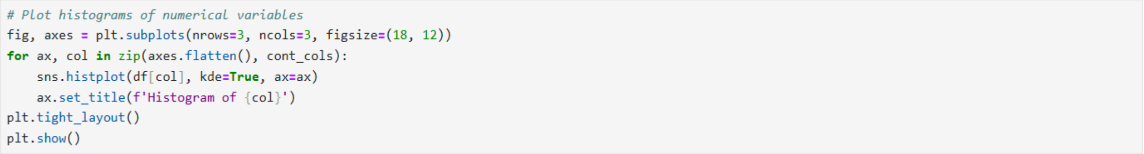
```python
# Check for Null Values
df.isnull().sum()
```



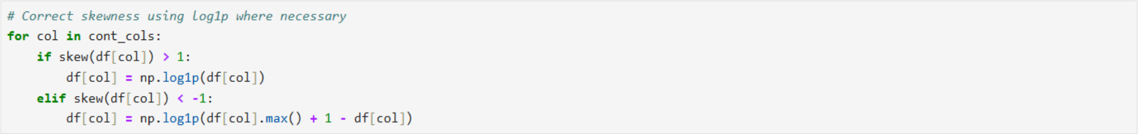
```python
# Drop Null Values
df.dropna(inplace=True)
```



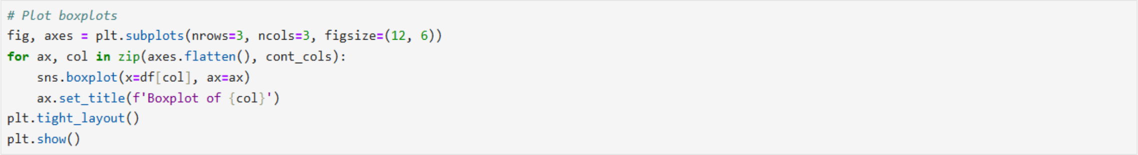
```python
# numerical columns
cont_cols = ['age', 'study_hours_per_day', 'social_media_hours', 'netflix_hours',
             'attendance_percentage', 'sleep_hours', 'exercise_frequency',
             'mental_health_rating', 'exam_score']
```



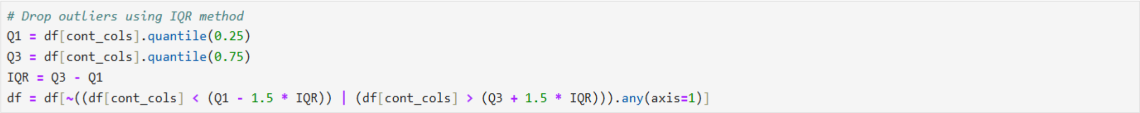
```python
# Plot histograms of numerical variables
fig, axes = plt.subplots(nrows=3, ncols=3, figsize=(18, 12))
for ax, col in zip(axes.flatten(), cont_cols):
    sns.histplot(df[col], kde=True, ax=ax)
    ax.set_title(f'Histogram of {col}')
plt.tight_layout()
plt.show()
```



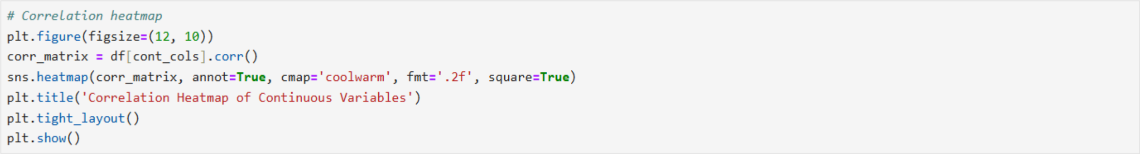
```
# Correct skewness using log1p where necessary
for col in cont_cols:
    if skew(df[col]) > 1:
        df[col] = np.log1p(df[col])
    elif skew(df[col]) < -1:
        df[col] = np.log1p(df[col].max() + 1 - df[col])
```



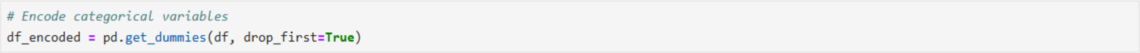
```
# Plot boxplots
fig, axes = plt.subplots(nrows=3, ncols=3, figsize=(12, 6))
for ax, col in zip(axes.flatten(), cont_cols):
    sns.boxplot(x=df[col], ax=ax)
    ax.set_title(f'Boxplot of {col}')
plt.tight_layout()
plt.show()
```



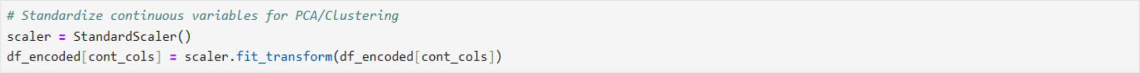
```
# Drop outliers using IQR method
Q1 = df[cont_cols].quantile(0.25)
Q3 = df[cont_cols].quantile(0.75)
IQR = Q3 - Q1
df = df[~((df[cont_cols] < (Q1 - 1.5 * IQR)) | (df[cont_cols] > (Q3 + 1.5 * IQR))).any(axis=1)]
```



```
# Correlation heatmap
plt.figure(figsize=(12, 10))
corr_matrix = df[cont_cols].corr()
sns.heatmap(corr_matrix, annot=True, cmap='coolwarm', fmt='.2f', square=True)
plt.title('Correlation Heatmap of Continuous Variables')
plt.tight_layout()
plt.show()
```

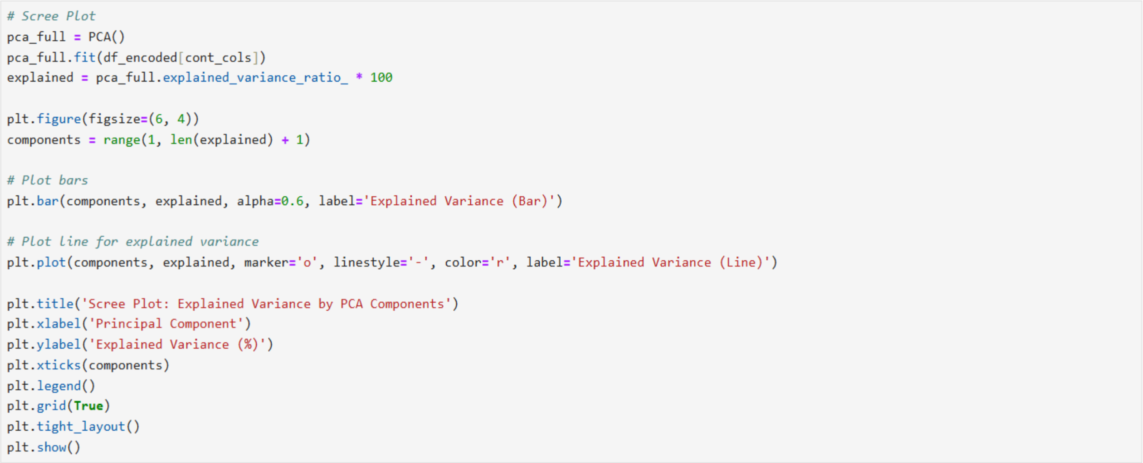


```
# Encode categorical variables
df_encoded = pd.get_dummies(df, drop_first=True)
```

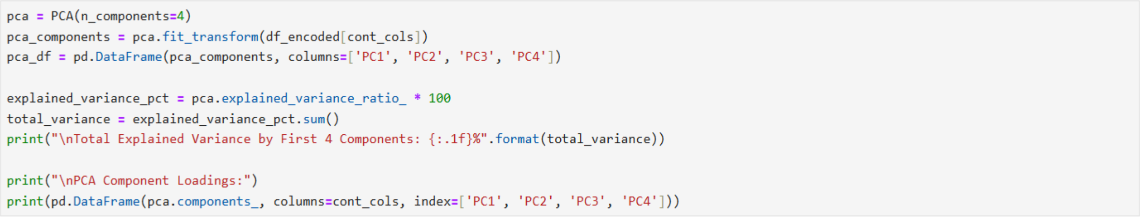


```
# Standardize continuous variables for PCA/Clustering
scaler = StandardScaler()
df_encoded[cont_cols] = scaler.fit_transform(df_encoded[cont_cols])
```

Principal Component Analysis (PCA)

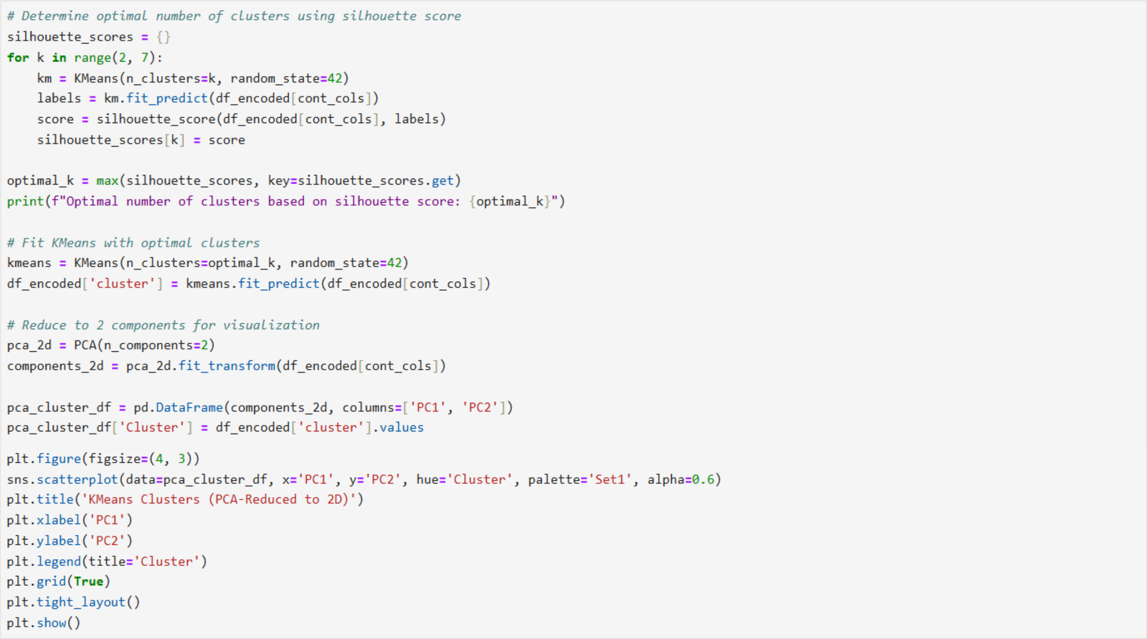


```
# Scree Plot
pca_full = PCA()
pca_full.fit(df_encoded[cont_cols])
explained = pca_full.explained_variance_ratio_ * 100

plt.figure(figsize=(6, 4))
components = range(1, len(explained) + 1)

# Plot bars
plt.bar(components, explained, alpha=0.6, label='Explained Variance (Bar)')

# Plot line for explained variance
plt.plot(components, explained, marker='o', linestyle='-', color='r', label='Explained Variance (Line)')

plt.title('Scree Plot: Explained Variance by PCA Components')
plt.xlabel('Principal Component')
plt.ylabel('Explained Variance (%)')
plt.xticks(components)
plt.legend()
plt.grid(True)
plt.tight_layout()
plt.show()
```



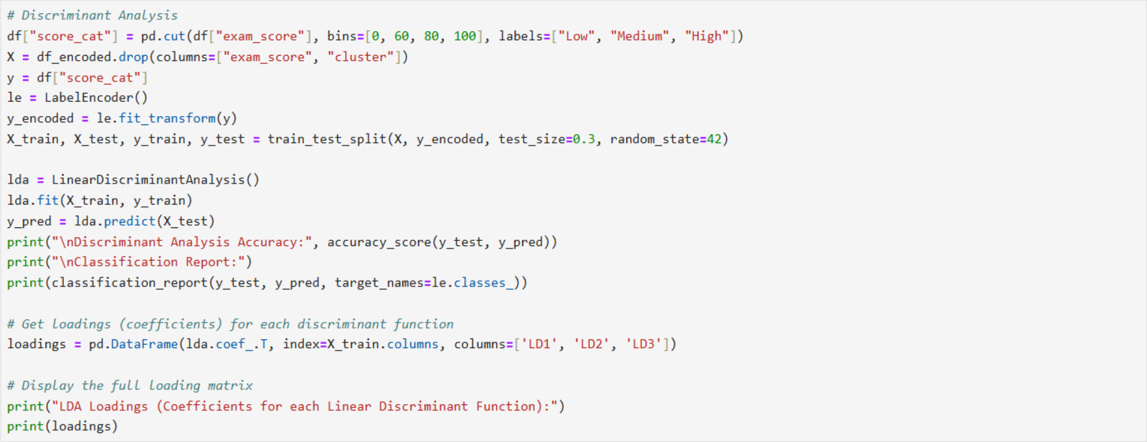
```
pca = PCA(n_components=4)
pca_components = pca.fit_transform(df_encoded[cont_cols])
pca_df = pd.DataFrame(pca_components, columns=['PC1', 'PC2', 'PC3', 'PC4'])

explained_variance_pct = pca.explained_variance_ratio_ * 100
total_variance = explained_variance_pct.sum()
print("\nTotal Explained Variance by First 4 Components: {:.1f}%".format(total_variance))

print("\nPCA Component Loadings:")
print(pd.DataFrame(pca.components_, columns=cont_cols, index=['PC1', 'PC2', 'PC3', 'PC4']))
```

Cluster Analysis (K-Means)



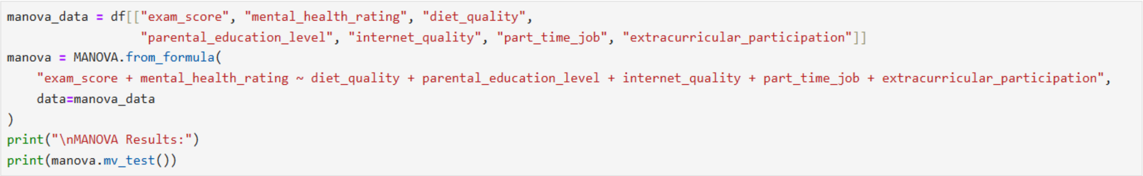
```python
# Determine optimal number of clusters using silhouette score
silhouette_scores = {}
for k in range(2, 7):
    km = KMeans(n_clusters=k, random_state=42)
    labels = km.fit_predict(df_encoded[cont_cols])
    score = silhouette_score(df_encoded[cont_cols], labels)
    silhouette_scores[k] = score

optimal_k = max(silhouette_scores, key=silhouette_scores.get)
print(f"Optimal number of clusters based on silhouette score: {optimal_k}")

# Fit KMeans with optimal clusters
kmeans = KMeans(n_clusters=optimal_k, random_state=42)
df_encoded['cluster'] = kmeans.fit_predict(df_encoded[cont_cols])

# Reduce to 2 components for visualization
pca_2d = PCA(n_components=2)
components_2d = pca_2d.fit_transform(df_encoded[cont_cols])

pca_cluster_df = pd.DataFrame(components_2d, columns=['PC1', 'PC2'])
pca_cluster_df['Cluster'] = df_encoded['cluster'].values

plt.figure(figsize=(4, 3))
sns.scatterplot(data=pca_cluster_df, x='PC1', y='PC2', hue='Cluster', palette='Set1', alpha=0.6)
plt.title('KMeans Clusters (PCA-Reduced to 2D)')
plt.xlabel('PC1')
plt.ylabel('PC2')
plt.legend(title='Cluster')
plt.grid(True)
plt.tight_layout()
plt.show()
```

Discriminant Analysis (LDA)



```python
# Discriminant Analysis
df["score_cat"] = pd.cut(df["exam_score"], bins=[0, 60, 80, 100], labels=["Low", "Medium", "High"])
X = df_encoded.drop(columns=["exam_score", "cluster"])
y = df["score_cat"]
le = LabelEncoder()
y_encoded = le.fit_transform(y)
X_train, X_test, y_train, y_test = train_test_split(X, y_encoded, test_size=0.3, random_state=42)

lda = LinearDiscriminantAnalysis()
lda.fit(X_train, y_train)
y_pred = lda.predict(X_test)
print("\nDiscriminant Analysis Accuracy:", accuracy_score(y_test, y_pred))
print("\nClassification Report:")
print(classification_report(y_test, y_pred, target_names=le.classes_))

# Get loadings (coefficients) for each discriminant function
loadings = pd.DataFrame(lda.coef_.T, index=X_train.columns, columns=['LD1', 'LD2', 'LD3'])

# Display the full loading matrix
print("LDA Loadings (Coefficients for each Linear Discriminant Function):")
print(loadings)
```

MANOVA



```python
manova_data = df[["exam_score", "mental_health_rating", "diet_quality",
                  "parental_education_level", "internet_quality", "part_time_job", "extracurricular_participation"]]
manova = MANOVA.from_formula(
    "exam_score + mental_health_rating ~ diet_quality + parental_education_level + internet_quality + part_time_job + extracurricular_participation",
    data=manova_data
)
print("\nMANOVA Results:")
print(manova.mv_test())
```