Mental Health Data Analysis: A Multivariate Statistical Approach

R.Pridhurshika

Registration No: S/19/484

Department of Statistics and Computer Science
Faculty of Science
University of Peradeniya

May 2025

Contents

Introduction

Mental health disorders have become a growing global concern, with factors such as stress, sleep habits, work-life balance, and physical activity commonly identified as key contributors. Gaining insight into how demographic and lifestyle characteristics influence mental health outcomes is essential for informing effective public health strategies.

1.1 Objectives:

- This study aims to explore the multivariate relationships among demographic characteristics, lifestyle factors, and mental health outcomes using several multivariate techniques.
- The goal is to determine whether these variables can meaningfully explain or predict the presence and severity of mental health conditions.

Methodology

2.1 Data Description

The dataset comprises 1000 anonymized observations collected across countries like the USA, India, the UK, Canada, and Australia. Each row represents an individual with the following types of data:

Variable	Type	Description
Age	Numeric	Age of the participant
Gender	Categorical	Self-identified gender
Occupation	Categorical	Participant's profession
Country	Categorical	Country of residence
Mental_Health_Condition	Categorical	Whether a mental health condition is reported
Severity	Categorical	Severity of condition
Consultation_History	Categorical	Whether the participant has consulted a mental health professional
Stress_Level	Categorical	Perceived stress level
Sleep_Hours	Numeric	Average hours of sleep per day
Work_Hours	Numeric	Average weekly work hours
Physical_Activity_Hours	Numeric	Average weekly physical activity hours

Table 1: Description of the variables used in the mental health dataset

2.2 Data Preprocessing

To ensure the dataset was suitable for multivariate statistical analysis, several preprocessing steps were carried out:

- **Missing Values:** Checked for and confirmed that there are no missing values in the dataset using `colSums(is.na(...))`.
- **Data Type Conversion:** Categorical variables such as `Gender`, `Occupation`, `Country`, `Mental_Health_Condition`, `Consultation_History`, and `Severity` were converted into factors for appropriate modeling.
- **Numeric Variable Scaling:** Variables used in multivariate techniques (CCA, PCA, SEM) such as `Age`, `Sleep_Hours`, `Work_Hours`, and `Physical_Activity_Hours` were standardized using the `scale()` function.
- **Derived Variables:** For Canonical Correlation Analysis and SEM, numeric representations of categorical variables were created using `as.numeric()` and appropriate factor level assignments to allow modeling in packages requiring numeric input.
- **Outlier Detection:** Boxplots were used to visually inspect numeric variables for potential outliers.

- **Visualization:** Histograms and pie charts were created to understand the distributions of key variables.

2.3 Statistical Techniques Used

The following statistical techniques were employed to explore and model the relationships within the mental health dataset:

- **Descriptive Statistics:** Used to summarize and understand the distribution and patterns of each variable.
- **Canonical Correlation Analysis (CCA):** Applied to assess the multivariate relationships between sets of demographic/lifestyle variables and mental health indicators.
- **Principal Component Analysis(PCA):** Used to reduce dimensionality and identify latent structures among variables.
- **Factor Analysis(FA):**To identify latent structures among numeric variables
- **Structural Equation Modeling (SEM):** Implemented To test hypothesized relationships among latent variables.
- **Discriminant Analysis:** Employed to classify individuals based on the presence or absence of a mental health condition, and to identify the most influential predictors.

All statistical analyses were performed using the R programming language (version 4.3.3), with the help of packages including `ggplot2`, `dplyr`, `CCA`, `psych`, `lavaan`, and `caret`.

Results and Discussion

This section discusses the results of each statistical method used to examine the relationships between demographic, lifestyle, and mental health indicators. Each method was selected to address specific analytical goals, and the results are interpreted in light of those goals.

3.1 Descriptive Analysis

Descriptive analysis provided an overview of the dataset. The age of the participants ranged from 18 to 65 years, with a mean of approximately 41.9 years. Participants reported an average of 7.1 hours of sleep per day, 54.6 hours of work per week, and 5.1 hours of physical activity per week. The gender distribution was balanced between categories and the participants came from various occupational and geographic backgrounds. These statistics suggest a well-rounded and diverse dataset suitable for multivariate exploration. Insert PCA plot and interpretation here.

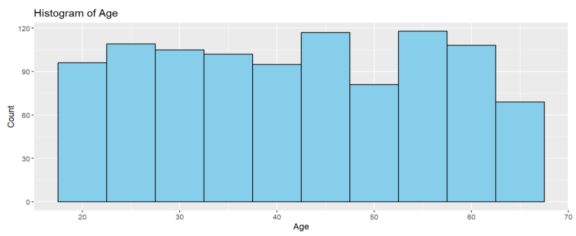


Figure 1: Histogram of Age

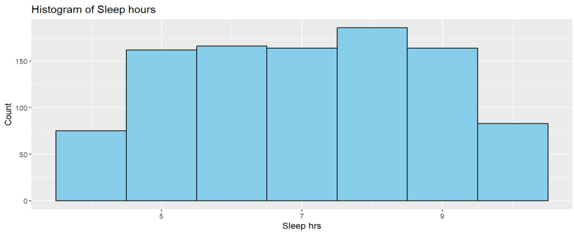


Figure 2: Histogram of Sleep hours

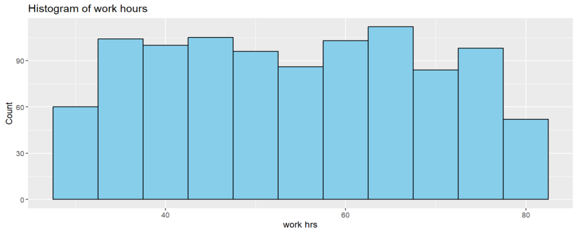


Figure 3: Histogram of Work hours

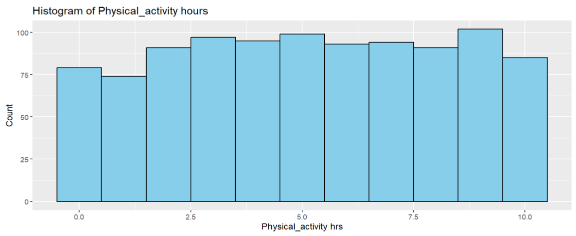
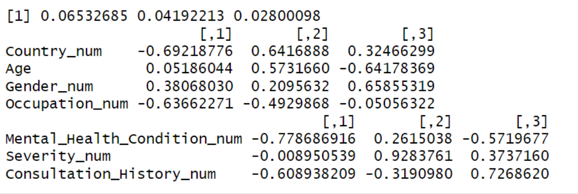


Figure 4: Histogram of Physical activity hours

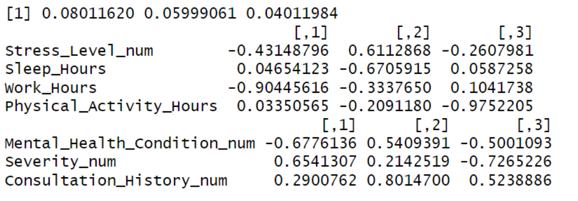
3.2 Canonical Correlation Analysis

CCA identifies the relationships between two sets of multivariate variables by finding linear combinations (canonical variates) that are maximally correlated. Here, CCA was used to explore the relationships between sets of variables: [1] *demographic versus mental health indicator* and [2] *lifestyle predictors versus mental health indicators.*



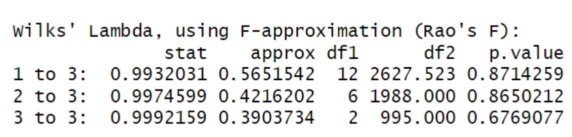
```
[1] 0.06532685 0.04192213 0.02800098
                    [,1]       [,2]        [,3]
Country_num    -0.69218776  0.6416888  0.32466299
Age             0.05186044  0.5731660 -0.64178369
Gender_num      0.38068030  0.2095632  0.65855319
Occupation_num -0.63662271 -0.4929868 -0.05056322
                                 [,1]       [,2]       [,3]
Mental_Health_Condition_num -0.778686916  0.2615038 -0.5719677
Severity_num                -0.008950539  0.9283761  0.3737160
Consultation_History_num    -0.608938209 -0.3190980  0.7268620
```

Figure 5: CCA result of [1]



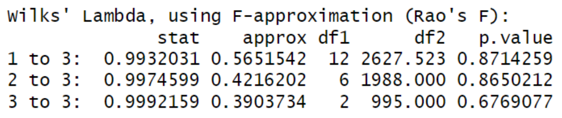
```
[1] 0.08011620 0.05999061 0.04011984
                              [,1]       [,2]       [,3]
Stress_Level_num          -0.43148796  0.6112868 -0.2607981
Sleep_Hours                0.04654123 -0.6705915  0.0587258
Work_Hours                -0.90445616 -0.3337650  0.1041738
Physical_Activity_Hours    0.03350565 -0.2091180 -0.9752205
                                 [,1]      [,2]       [,3]
Mental_Health_Condition_num -0.6776136 0.5409391 -0.5001093
Severity_num                 0.6541307 0.2142519 -0.7265226
Consultation_History_num     0.2900762 0.8014700  0.5238886
```

Figure 6: CCA result of [2]



```
Wilks' Lambda, using F-approximation (Rao's F):
             stat    approx df1      df2   p.value
1 to 3:  0.9932031 0.5651542  12 2627.523 0.8714259
2 to 3:  0.9974599 0.4216202   6 1988.000 0.8650212
3 to 3:  0.9992159 0.3903734   2  995.000 0.6769077
```

Figure 7: Wilks' Lambda Test result of [1]



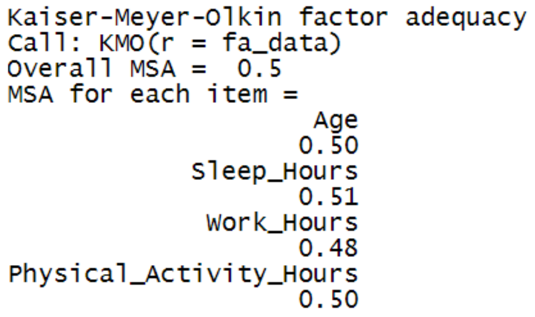
```
Wilks' Lambda, using F-approximation (Rao's F):
             stat    approx df1      df2   p.value
1 to 3:  0.9932031 0.5651542  12 2627.523 0.8714259
2 to 3:  0.9974599 0.4216202   6 1988.000 0.8650212
3 to 3:  0.9992159 0.3903734   2  995.000 0.6769077
```

Figure 8: Wilks' Lambda Test result of [2]

In both cases the canonical correlations were low and statistically insignificant ($p-values > 0.8$). This result implies that neither demographic factors nor lifestyle factors had a strong multivariate association with mental health outcomes such as reported conditions, severity, or consultation history.

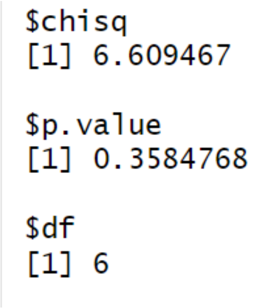
3.3 Factor Analysis(FA)

FA is used to identify underlying latent variables that explain the pattern of correlations among observed variables. It is especially useful for data reduction and exploring the dimensional structure of a dataset. Here, Factor analysis was considered to identify latent variables among numeric variables.



```
Kaiser-Meyer-Olkin factor adequacy
Call: KMO(r = fa_data)
Overall MSA =  0.5
MSA for each item = 
                    Age 
                   0.50 
            Sleep_Hours 
                   0.51 
             Work_Hours 
                   0.48 
Physical_Activity_Hours 
                   0.50 
```

Figure 9: Result of KMO test



```
$chisq
[1] 6.609467

$p.value
[1] 0.3584768

$df
[1] 6
```

Figure 10: Result of Bartlett's Test

However, diagnostic tests indicated that this approach was not suitable. The Kaiser-Meyer-Olkin (KMO) value was exactly 0.5, which is the minimum acceptable threshold, and Bartlett's test yielded a non-significant p-value (0.3585), indicating weak correlations among variables. Consequently, no meaningful latent structure could be extracted from the data and factor analysis was not pursued further.

3.4 Principal Component Analysis

PCA reduces the dimensionality of a dataset by transforming variables into a smaller number of uncorrelated principal components that retain most of the original variance.Since FA was not pursued in this study, PCA was used to examine variance patterns and reduce dimensionality among continuous variables.

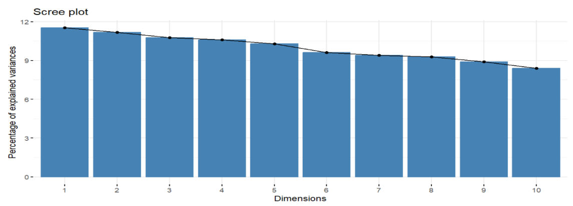
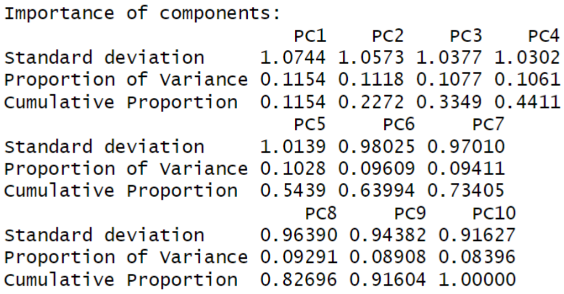


Figure 11: Sree Plot



```
Importance of components:
                          PC1    PC2    PC3    PC4
Standard deviation     1.0744 1.0573 1.0377 1.0302
Proportion of Variance 0.1154 0.1118 0.1077 0.1061
Cumulative Proportion  0.1154 0.2272 0.3349 0.4411
                          PC5     PC6     PC7
Standard deviation     1.0139 0.98025 0.97010
Proportion of Variance 0.1028 0.09609 0.09411
Cumulative Proportion  0.5439 0.63994 0.73405
                           PC8     PC9    PC10
Standard deviation     0.96390 0.94382 0.91627
Proportion of Variance 0.09291 0.08908 0.08396
Cumulative Proportion  0.82696 0.91604 1.00000
```

Figure 12: Result of PCA

The PCA output shows that the first five principal components explain about 54.4% of the total variance, with each component contributing roughly 10–11%. The variance is evenly distributed, indicating no dominant component. To retain over 75% of the variance, at least seven components are needed. This suggests that dimensionality reduction is possible but would require multiple components to preserve most of the information.

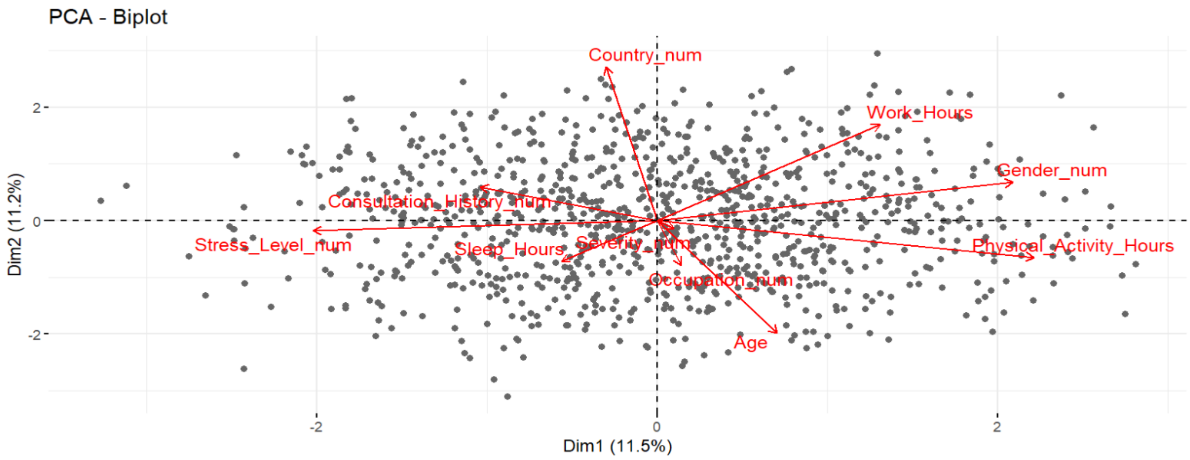


Figure 13: BiPlot

3.5 Structure Equation Model(SEM)

SEM is a multivariate statistical technique for testing hypotheses about relationships among observed and latent variables. It combines factor analysis and multiple regression in one framework.Here, SEM was used to assess how demographic and lifestyle variables jointly influence mental health outcomes through latent constructs.

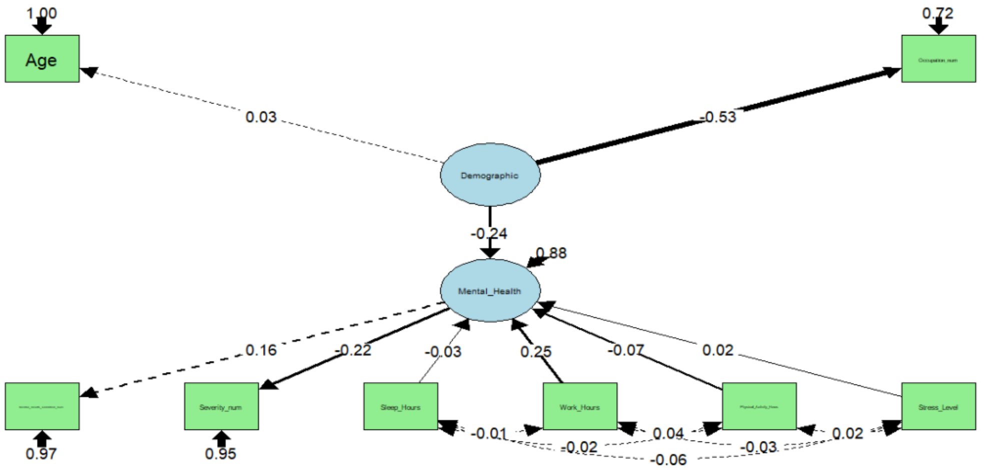
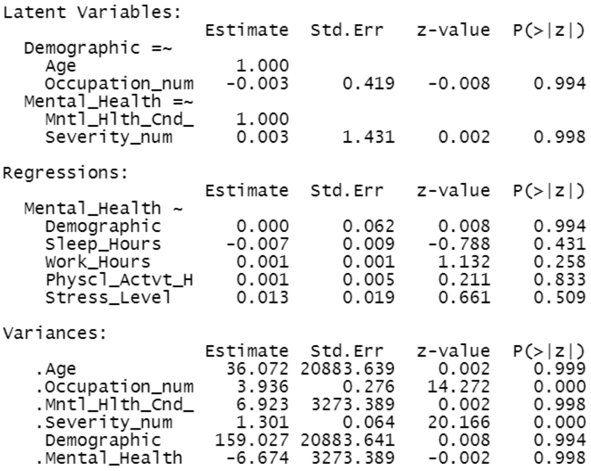


Figure 14: SEM plot



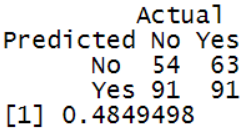
```
Latent Variables:
                   Estimate  Std.Err  z-value  P(>|z|)
  Demographic =~
    Age               1.000
    Occupation_num   -0.003    0.419   -0.008    0.994
  Mental_Health =~
    Mntl_Hlth_Cnd_    1.000
    Severity_num      0.003    1.431    0.002    0.998

Regressions:
                   Estimate  Std.Err  z-value  P(>|z|)
  Mental_Health ~
    Demographic       0.000    0.062    0.008    0.994
    Sleep_Hours      -0.007    0.009   -0.788    0.431
    Work_Hours        0.001    0.001    1.132    0.258
    Physcl_Actvt_H    0.001    0.005    0.211    0.833
    Stress_Level      0.013    0.019    0.661    0.509

Variances:
                   Estimate  Std.Err  z-value  P(>|z|)
   .Age              36.072 20883.639    0.002    0.999
   .Occupation_num    3.936    0.276   14.272    0.000
   .Mntl_Hlth_Cnd_    6.923 3273.389    0.002    0.998
   .Severity_num      1.301    0.064   20.166    0.000
    Demographic     159.027 20883.641    0.008    0.994
   .Mental_Health    -6.674 3273.389   -0.002    0.998
```

Figure 15: SEM results

The model demonstrated excellent fit indices (CFI = 1.0, RMSEA = 0.000, SRMR = 0.015), confirming that the model structure was appropriate. However, none of the paths from demographics or lifestyle variables to mental health indicators were statistically significant. This further supports earlier findings that these observed variables alone are insufficient to explain mental health variability in this dataset.

3.6 Discriminant Analysis

LDA was performed to classify individuals into categories of mental health condition (yes / no) using demographic and lifestyle predictors Three types of discriminant analysis were used to classify individuals based on their mental health status:Linear Discriminant Analysis (LDA), Quadratic Discriminant Analysis (QDA), and Regularized Discriminant Analysis (RDA).



```
          Actual
Predicted No Yes
      No  54  63
      Yes 91  91
[1] 0.4849498
```

Figure 16: LDA Classification Plot

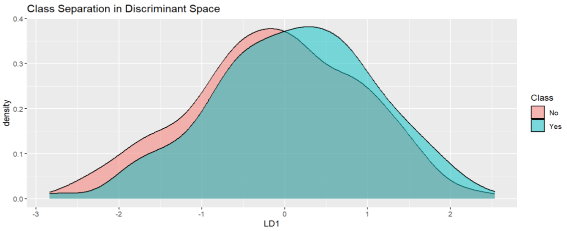
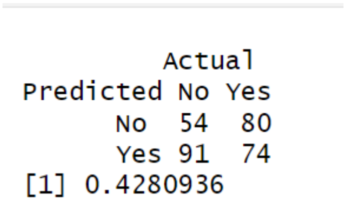


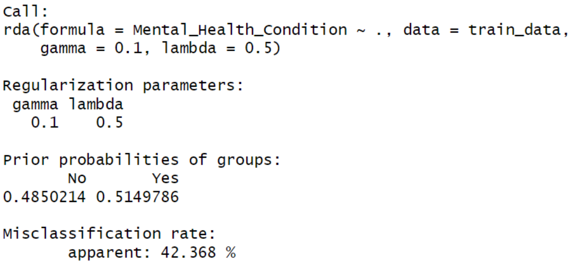
Figure 17: Class separation in discriminant space

The performance of all three models was suboptimal. LDA achieved an accuracy of only 48.5% which is marginally above the baseline of random classification.The plot shows that although some separation exists between the classes along LD1, it is not strong enough to clearly classify individuals based on the available variables. This supports your finding that demographic and lifestyle features in this dataset are limited in their ability to predict mental health conditions effectively using LDA.



```
          Actual
Predicted No Yes
      No  54  80
      Yes 91  74
[1] 0.4280936
```

Figure 18: QDA classification plot



```
Call:
rda(formula = Mental_Health_Condition ~ ., data = train_data, 
    gamma = 0.1, lambda = 0.5)

Regularization parameters:
 gamma lambda 
   0.1    0.5 

Prior probabilities of groups:
       No       Yes 
0.4850214 0.5149786 

Misclassification rate:
       apparent: 42.368 %
```

Figure 19: RDA results

QDA performed even worse, with an accuracy of 42.8%. RDA, which incorporates regularization to handle multicollinearity, resulted in an error rate of 51.5%, indicating no substantial improvement.

Conclusion

This study applied a comprehensive set of multivariate statistical techniques to explore the relationships between demographic characteristics, lifestyle behaviors, and mental health outcomes. Despite the diversity and quality of the dataset, the results across all analytical methods consistently indicated weak or non-significant associations.

Canonical Correlation Analysis, Factor Analysis, and Principal Component Analysis did not reveal any strong underlying structures or meaningful multivariate relationships. Although the Structural Equation Model demonstrated excellent fit indices, the hypothesized paths from demographic and lifestyle factors to mental health outcomes were not statistically significant. Furthermore, classification techniques such as Linear Discriminant Analysis (LDA), Quadratic Discriminant Analysis (QDA), and Regularized Discriminant Analysis (RDA) performed poorly, with classification accuracies barely exceeding random chance.

These findings suggest that commonly used demographic and lifestyle variables may be insufficient on their own to explain or predict mental health conditions.

Recommendation

- Future research should incorporate validated mental health instruments, such as the PHQ-9 for depression and GAD-7 for anxiety, to enhance the accuracy and reliability of psychological assessments.
- A longitudinal study design is recommended to better understand changes in mental health over time and to facilitate causal inferences.
- The application of machine learning techniques could help identify complex, non-linear relationships and interaction effects that may be overlooked by traditional statistical methods.

References

- Husson, F., Lê, S., & Pagès, J. (2011). Exploratory multivariate analysis by example using R (Vol. 15, pp. 1-60). Boca Raton: CRC press .
- Everitt, B., & Hothorn, T. (2011). An introduction to applied multivariate analysis with R. Springer Science & Business Media.

- Ramayah, T., Ahmad, N. H., Halim, H. A., & May-Chiun, S. R. M. Z. (2010). Discriminant analysis: An illustrated example. African Journal of Business Management, 4(9), 1654.

- https://lavaan.ugent.be/

Appendices

Appendix A: Dataset Snapshot

Age	Gender	Country	Occupation	Sleep	Stress	WorkHours	PhysicalActivity
25	Male	USA	Engineer	7	Medium	40	3
...	...	...	...	...	...	...	...

(Only a sample. Full dataset available on Sleep Health and Lifestyle Dataset)

Appendix B: R Code (Markdown Format)

Canonical Correlation Analysis (CCA)

```r
library(CCA)
library(CCP)

X <- mental_health_data %>% select(Age, Sleep_Hours, Exercise_Frequency)
Y <- mental_health_data %>% select(Stress_Level, Anxiety_Level, Depression_Score)

sum(is.na(X))
sum(is.na(Y))

cca_results <- cc(X, Y)
cca_results$cor
```

```

rho <- cca_results$cor
n <- nrow(X)
p <- ncol(X)
q <- ncol(Y)

wilks_stat <- prod(1 - rho^2)
chi_sq <- -((n - 1) - (p + q + 1) / 2) * log(wilks_stat)
df <- p * q
p_value <- pchisq(chi_sq, df = df, lower.tail = FALSE)

cat("Wilks' Lambda:", wilks_stat, "\n")
cat("Chi-squared statistic:", chi_sq, "\n")
cat("Degrees of freedom:", df, "\n")
cat("P-value:", p_value, "\n")
```

Listing 1: CCA Code

Principal Component Analysis (PCA)

```
library(FactoMineR)
library(factoextra)

pca_vars <- mental_health_data %>%
  select(Age, Sleep_Hours, Exercise_Frequency, Stress_Level, Anxiety_Level, Depression_Score)

pca_vars_scaled <- scale(pca_vars)

pca_result <- PCA(pca_vars_scaled, graph = FALSE)

fviz_eig(pca_result, addlabels = TRUE, ylim = c(0, 50))
fviz_pca_ind(pca_result, col.ind = "cos2", gradient.cols = c("blue", "green", "red"), repel = TRUE)
```

```
fviz_pca_var(pca_result, col.var = "contrib", gradient.cols = c("blue", "green", "red"), repel = TRUE)
```

Listing 2: PCA Code

Factor Analysis (FA)

```
library(psych)

fa_vars <- mental_health_data %>%
  select(Age, Sleep_Hours, Exercise_Frequency, Stress_Level, Anxiety_Level, Depression_Score)

KMO_result <- KMO(fa_vars)
print(KMO_result)

cortest.bartlett(fa_vars)

fa_result <- fa(fa_vars, nfactors = 2, rotate = "varimax", fm = "ml")
print(fa_result)
```

Listing 3: FA Code

Structural Equation Modeling (SEM)

```
library(lavaan)

sem_model <- '
  # Measurement model
  Mental_Health =~ Stress_Level + Anxiety_Level + Depression_Score
  Lifestyle =~ Age + Sleep_Hours + Exercise_Frequency

  # Structural model
  Mental_Health ~ Lifestyle
```

```r
'

fit <- sem(sem_model, data = mental_health_data, missing = "fiml")

summary(fit, fit.measures = TRUE, standardized = TRUE)
```

Listing 4: SEM Code

Discriminant Analysis (LDA)

```r
library(MASS)

# Assume mental_health_data has a factor variable 'Group' (e.g., low, medium, high stress)
lda_data <- mental_health_data %>%
  select(Group, Age, Sleep_Hours, Exercise_Frequency, Stress_Level, Anxiety_Level, Depression_Score)

lda_result <- lda(Group ~ ., data = lda_data)

print(lda_result)
plot(lda_result)

predictions <- predict(lda_result)
table(lda_data$Group, predictions$class)

# Accuracy
mean(lda_data$Group == predictions$class)
```

Listing 5: Discriminant Analysis (LDA) Code