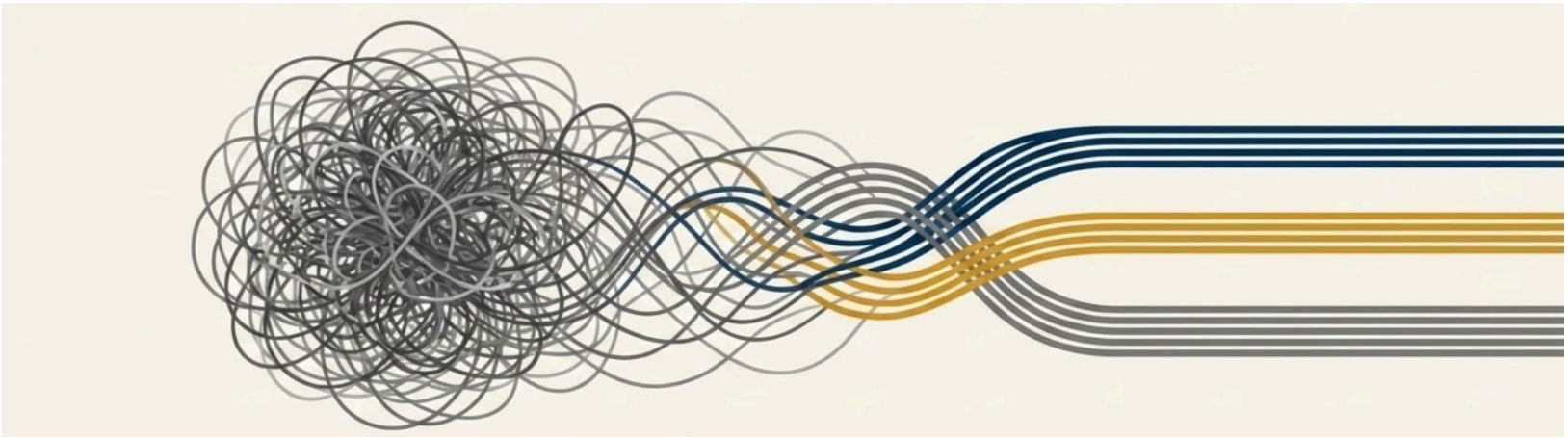
Untangling Student Stress:

A Multivariate Analysis Perspective



Group 2: Guus Bouwens, Clay Cleavinger, Ty Campbell, Chris Kerr

1. Introduction and Research Framework

The mental health of university students is an issue of critical importance, with elevated stress levels consistently linked to diminished academic performance and overall well-being (Andersen et al., 2021). While many factors contributing to student stress are studied, they are often examined in isolation. This approach fails to capture the complex, interconnected nature of psychological, environmental, and academic pressures. A holistic understanding requires a multivariate analytical framework capable of dissecting these intricate relationships simultaneously (Spagert et al., 2022). This report presents a case study in methodological triangulation, detailing the statistical techniques employed to deconstruct the "Student Stress Factors" dataset, moving beyond simple correlations to identify latent structures, segment the student population, and uncover specific, actionable insights.

The core research objectives of this methodological study were as follows:
• Dimension Reduction: To reduce the 20 observed variables into a smaller, more manageable set of uncorrelated principal components that capture the maximum possible variance in the data.
• Latent Structure Identification: To identify and model the unobserved latent constructs, such as "Psychological Distress," that explain the patterns of correlation among the observed variables.
• Structural Validation: To formally test and validate the hypothesized latent factor structure of student stress using Confirmatory Factor Analysis (CFA).
• Student Segmentation: To partition the student population into distinct, homogeneous groups or profiles based on their stress-related characteristics using a combination of clustering algorithms.
• Association Discovery: To uncover significant "if-then" relationships and predictive patterns between different stress factors using Market Basket Analysis.

The analysis was conducted on a dataset comprising survey responses from 1,100 university students. The data consisted of 20 numerical variables representing a range of psychological, academic, and social factors. The blood_pressure variable was

excluded from the final analysis due to preliminary findings indicating low communality and a poor fit with the remaining variables, suggesting it did not meaningfully contribute to the shared constructs being measured.

This report begins by outlining the foundational data preparation and diagnostic procedures that precede any robust multivariate analysis.

2. Data Preparation and Preliminary Diagnostics

Rigorous data preparation is a non-negotiable prerequisite for any multivariate analysis. Methodologies such as Principal Component Analysis (PCA) and Cluster Analysis are highly sensitive to the scale and distribution of the input variables, while methods like Factor Analysis rely on specific assumptions about the data's underlying structure. Therefore, a strategic sequence of standardization and diagnostic testing was performed to ensure the integrity and validity of all subsequent analytical steps.

The Data

Our dataset represents a variety of self-reported features which would be conducive to stress levels. These variables include psychological factors: anxiety level 0-30 scale, self esteem 0-30, history of mental health problems 0-1, depression level 0-30. Physiological factors: frequency of headaches 0-5, blood pressure levels (removed), sleep quality 0-5, breathing issues 0-5. Environmental factors: noise level when studying 0-5, living conditions 0-5, everyday safety 0-5, and if the student has their basic needs met 0-5. Academic factors: academic performance 0-5, study load 0-5, strength of teacher-student relationship 0-5, and outlook on future career opportunities 0-5. And finally social factors: social support from people around them 0-3, peer pressure 0-5, extracurricular activities 0-5, and frequency/level of bullying they experience 0-5. All of these variables seek to explain stress level, the final outcome variable, which could be high, medium, or low. As you can see, there are a variety of different scales and ratings associated with these variables. This variance is the motivation behind the standardization process, which will ensure that each variable is properly scaled and ready to be modeled using our multivariate analysis techniques.

Data Standardization

Standardization is a prerequisite for multivariate techniques that rely on variance or distance metrics. All 20 variables in the dataset were transformed into z-scores, yielding a distribution for each variable with a mean of 0 and a variance of 1. This step is crucial for removing the arbitrary influence of a variable's original scale. Without standardization, variables measured on a larger scale would disproportionately dominate the results of distance-based or variance-based algorithms when compared to variables measured on a smaller scale. Standardization ensures that each variable contributes to the analysis based on its correlation structure, not its measurement unit.

2.1 Assessment of Multivariate Normality

The assumption of multivariate normality was assessed using a Chi-square Q-Q plot of the squared Mahalanobis distances. This diagnostic tool plots the observed squared Mahalanobis distances against the corresponding quantiles of a Chi-square distribution. A noticeable deviation from the reference line was observed in the upper tail, indicating that the data likely violates the assumption of multivariate normality and may contain outliers. This departure from normality is a critical diagnostic finding that will necessitate a de-emphasis on Chi-Square-based fit statistics in favor of more robust indices during the subsequent factor analyses. Acknowledging this finding, the decision was made to proceed, supported by the large sample size (N=1,100), which can provide some robustness to such violations.

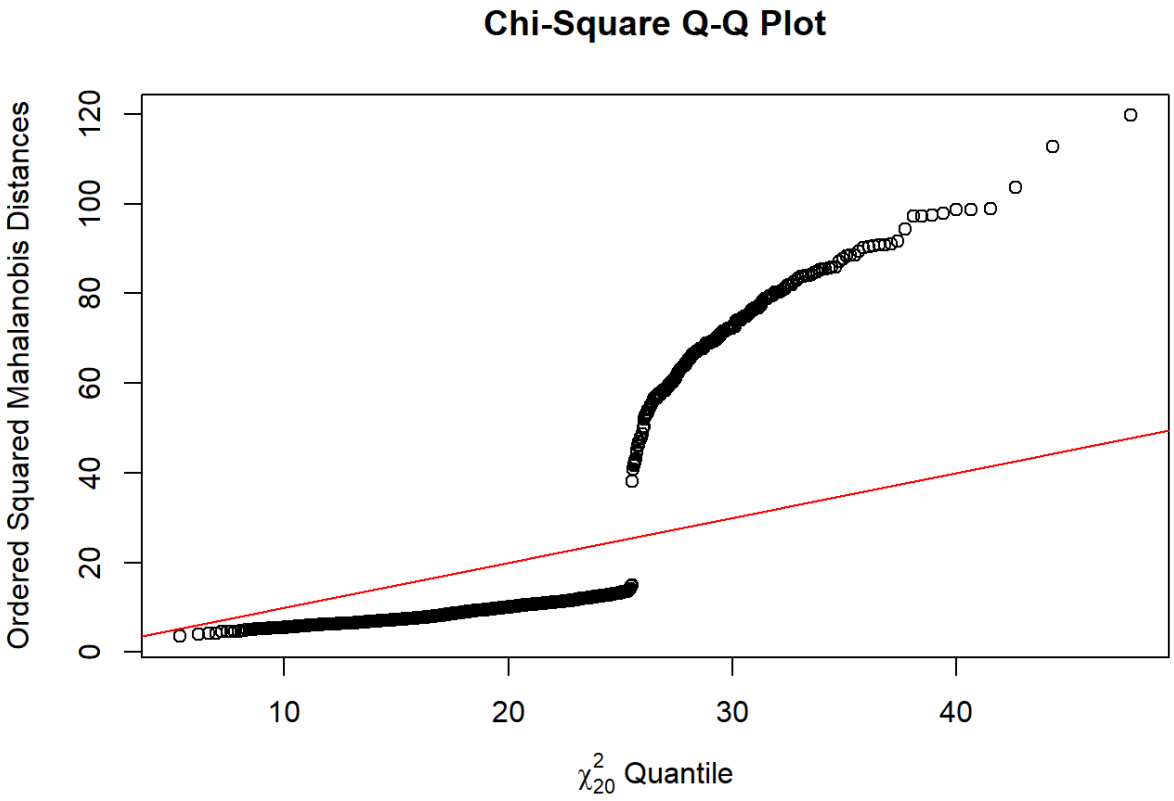


Figure 2.1.1: Chi^2 Q-Q Plot for student data. Appendix 2.2

As seen in the figure above, there is a clear deviation from multivariate normality. With that said however, there is vital information in the deviation. This is why we went forward with *standardization* through conversion to z-score rather than a *normalization* approach. For methods such as PCA, standard scales are required, not a normal distribution.

2.2 Correlation Analysis

A correlation matrix was generated and visualized as a heatmap to examine the inter-relationships among the 20 variables and determine the suitability of the dataset for dimension reduction. The correlation heatmap revealed a substantively meaningful structure, including a prominent cluster of inter-correlated psychological distress indicators (e.g., anxiety_level, depression, stress_level), confirming the data's suitability for factor analysis. This patterned structure strongly indicates that the variables are not

independent and share underlying common variance, making the dataset highly appropriate for techniques like PCA and Factor Analysis, which aim to model these shared structures.

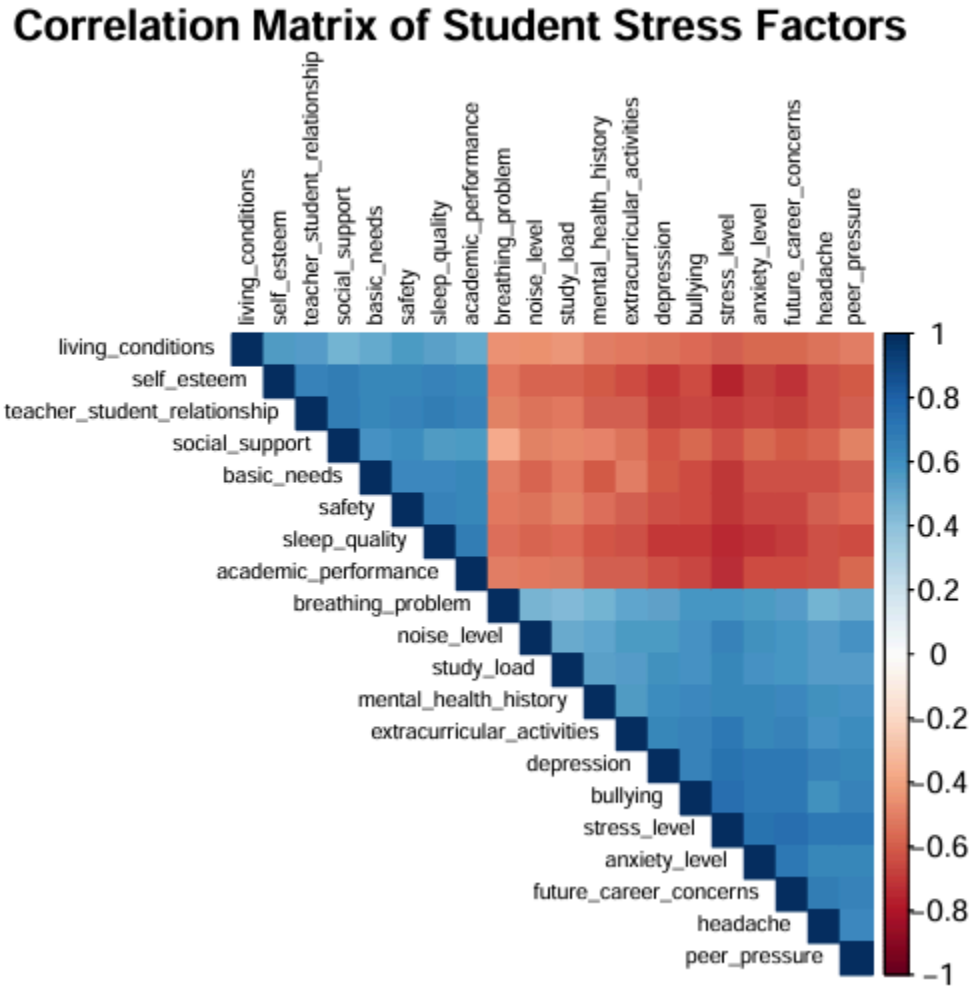


Figure 2.2.1: Correlation Matrix. Appendix 2.3

Having confirmed the suitability of the data, the analysis proceeded to the first primary analytical technique: dimension reduction.

3. Dimension Reduction and Latent Structure Analysis

This stage of the analysis involved a methodical progression from data simplification to theoretical model validation. The initial goal was to reduce the complexity of the 20-variable dataset using Principal Component Analysis (PCA). Subsequently, the

ambiguity in PCA's atheoretical components provided a clear mandate for the more theoretically-driven Exploratory and Confirmatory Factor Analysis (EFA/CFA) to identify, interpret, and validate the underlying constructs of student stress.

3.1 Principal Component Analysis (PCA)

The primary goal of applying PCA was to reduce the 20 observed variables into a smaller set of uncorrelated linear combinations, known as principal components, that collectively explain the maximum possible variance in the original data.

Determining the Number of Components

Multiple criteria guided the decision on component retention. While the Kaiser Criterion (eigenvalues > 1) is often used, its tendency to suggest too few or too many components is well-documented; in this case, it pointed to a single, overly broad component. The Scree Plot's inflection point, however, clearly suggested a 2- or 3-component solution. A 3-component solution was ultimately retained as it offered superior theoretical interpretability and thematic separation of the variance.

	PC1	PC2	PC3
anxiety_level	0.24	0.11	0.02
self_esteem	-0.24	0.17	0.08
mental_health_history	0.21	0.11	-0.11
depression	0.24	-0.03	-0.15
headache	0.22	-0.08	-0.15
sleep_quality	-0.24	-0.06	-0.04
breathing_problem	0.19	0.48	0.55
noise_level	0.20	0.18	-0.21
living_conditions	-0.19	0.01	-0.39
safety	-0.23	0.20	-0.29
basic_needs	-0.22	0.14	-0.09
academic_performance	-0.23	0.12	-0.14
study_load	0.20	0.19	-0.48
teacher_student_relationship	-0.23	0.30	-0.14
future_career_concerns	0.24	0.02	-0.03
social_support	-0.21	0.60	0.07
peer_pressure	0.22	0.25	-0.21
extracurricular_activities	0.22	0.19	-0.11
bullying	0.24	0.15	0.05
stress_level	0.25	0.03	-0.06

Figure 3.1.1: Principle Component Loadings, Appendix 3.1

Interpretation of Principal Components

The three retained principal components were interpreted by examining their loadings, which represent the correlation between each original variable and the component.

• PC1: General Distress. This component was defined by high positive loadings on variables such as anxiety_level, depression, and stress_level, and strong negative loadings on self_esteem and sleep_quality. It clearly represents a broad dimension of overall psychological distress.

• PC2: Social & Physiological Factors. This component was characterized by high positive loadings on social_support, teacher_student_relationship, and breathing_problem. This component appears to capture a mix of positive interpersonal connections and a somatic symptom.

• PC3: Environmental Factors. This component was defined by a high positive loading on breathing_problem and high negative loadings on study_load, noise_level, and living_conditions. This suggests a complex dimension where higher scores are associated with more breathing issues but a less demanding physical and academic

environment.

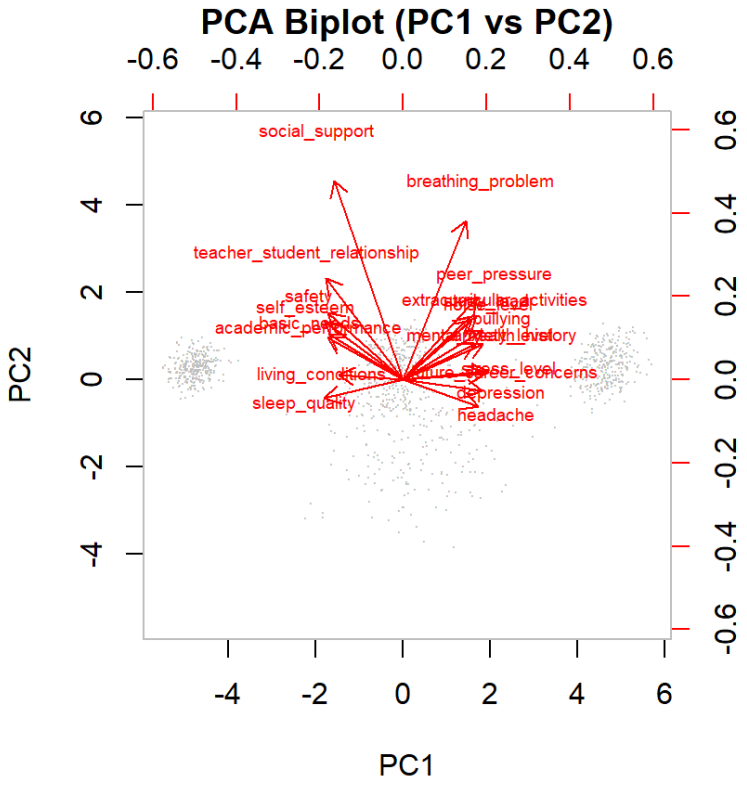


Figure 3.1.2: PCA Biplot of components 1 & 2, Appendix 3.4

The structure of the data in this reduced 3-dimensional space was further visualized using biplots and 3D plots, which map the observations and original variables onto the principal component axes.

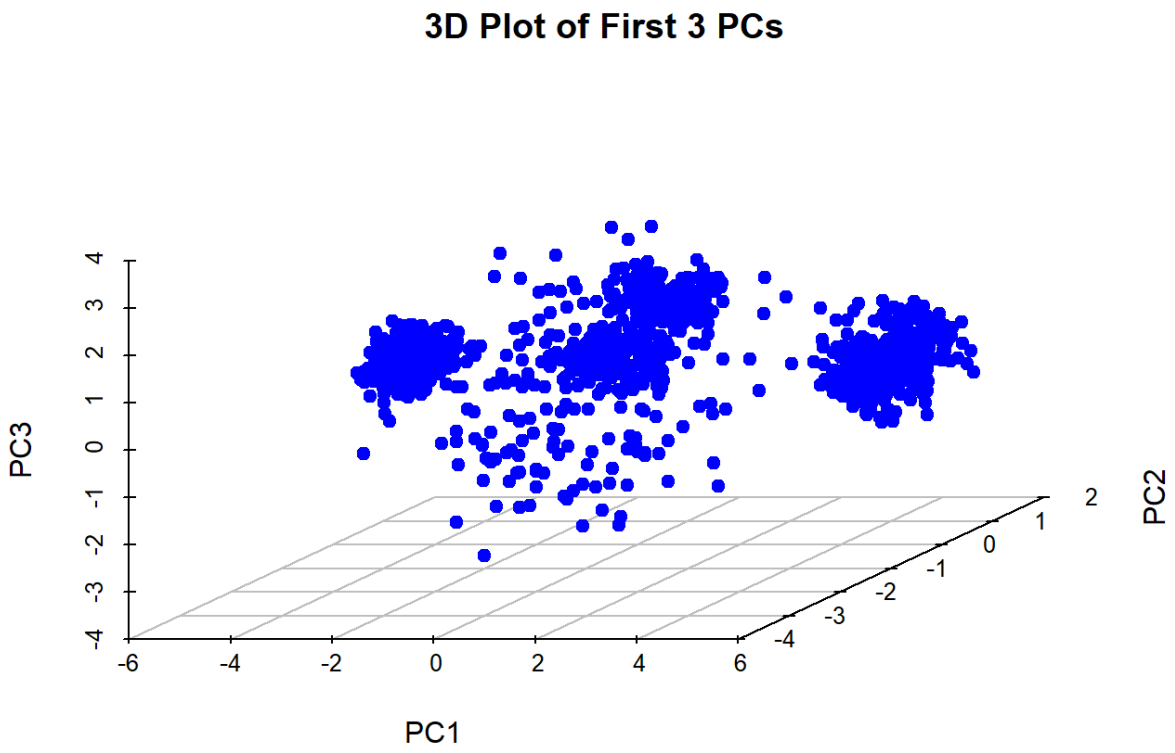


Figure 3.1.2: Perspective Plot of 1st 3 principle components. Appendix 3.5

Segments, 1-3.1, written by Chris Kerr.

3.2 Exploratory Factor Analysis (EFA)

While PCA provided a parsimonious 3-component solution, its components (particularly PC2 and PC3) were less theoretically "clean" than what is desired for construct validation. This motivated the shift to EFA, a statistical method aimed at uncovering the unobserved latent constructs, or factors, that are believed to cause the observed correlations among the variables.

Model Specification and Results

Guided by the PCA results, a 3-factor EFA model was specified and fitted using Maximum Likelihood estimation. A Promax (oblique) rotation was applied. This method

was selected over an orthogonal rotation (like Varimax) because it allows the latent factors to be correlated. This is a more theoretically sound assumption for psychological and behavioral data, where constructs like "Psychological Distress" and "Academic Pressure" are expected to be related rather than completely independent, and in this dataset are highly correlated.

Model Sufficiency Test

The Chi-square test for model sufficiency yielded a statistically significant result ($p < 0.05$), formally rejecting the null hypothesis that three factors are perfectly sufficient to explain the data. However, with a large sample size (N=1,100), the Chi-square test becomes extremely sensitive to even trivial deviations from a perfect fit. Therefore, greater weight was placed on the model's practical utility, strong theoretical interpretability, and the substantial proportion of variance explained by the three factors.

The results of our EFA is illustrated by the chart below. The chart below shows the loadings for the 3 factors for each variable in our data set.

	Factor 1	Factor 2	Factor 3
Anxiety Level	0.51		
Self Esteem	-0.54		
Mental Health History	0.46		
Depression	0.61		
Headache	0.47		
Sleep Quality		0.50	
Breathing Problem		-0.46	
Noise Level	0.56		

Living Conditions		0.38	
Safety		0.60	
Basic Needs		0.61	
Academic Performance		0.67	
Study Load	0.69		
Student Teacher Relationship		0.58	
Future Career Concerns	0.60		
Social Support			0.85
Peer Pressure	0.71		
Extracurricular Activities	0.76		
Bullying	0.47		
Stress Level	0.57		

Table 3.2.1: Factor Loadings of EFA analysis. Appendix 3.2

Factor 1 Loadings

Looking at the first factor, we can see that it is heavily loaded on all of our negative aspect variables (anxiety, stress level, depression, etc.) and negatively loaded on self esteem. This means that high scoring observations have high scores in all of those negative variables and low self esteem. They have high stress, a lot of concerns regarding their career, are in harsh environments, and just have a lot of chaos going on in their lives.

Factor 2 Loadings

For this second factor, it is heavily loaded on academic and environmental variables and then negatively loaded on breathing problems. This factor is likely indicating academic and environmental standings and how this contributes to anxiety, due to the breathing problem loading. High scoring observations in this factor are in good academic standing and are in good environments and have no breathing issues.

Factor 3 Loadings

Factor three only has one variable that it is loaded on, and that variable is social support. Because this is the only variable that it is loaded on, then we can infer that this factor represents the community surrounding the particular person. High scoring observations in this factor indicate individuals that have a strong community of people that is backing them up.

Factor Labeling

Because of our previous interpretations for each of the factors, we decided to provide specific label names for each factor. For the first factor, we decided to label it “Psychological Distress”. For the second factor, we decided to label it “Academic and Environmental Pressure”. As for the last factor, we decided to label it “Support & Environment”. These three latent variables that were identified will later be deeper analyzed in our next segment, CFA.

This segment (3.2) was written by Tyler Campbell

3.3 Confirmatory Factor Analysis (CFA)

Following the exploratory phase, Confirmatory Factor Analysis (CFA) was employed to formally test the hypothesized 3-factor measurement model derived from EFA. Its core question is: "Does the hypothesized 3-factor structure provide a good fit to the observed data?"

Hypothesized Model

The model specified for confirmation consisted of three latent factors, with observed variables assigned as indicators based on the EFA results and theoretical coherence:

• Psychological (Psy): Indicated by anxiety_level, self_esteem, depression, stress_level.

• Pressure (Prs): Indicated by peer_pressure, study_load, future_career_concerns.

• Support (Spp): Indicated by social_support, safety, basic_needs.

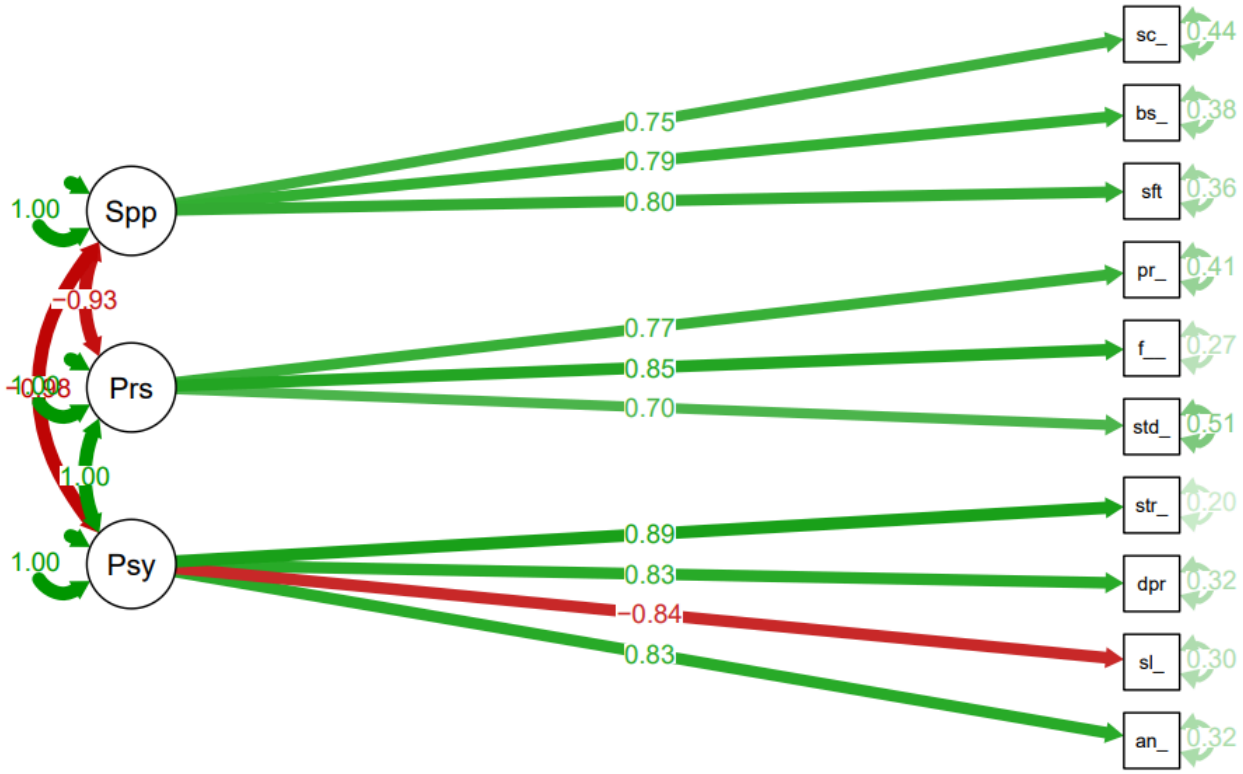


Figure 3.3.1: Graph of hypothesised CFA model with calculated path coefficients. Appendix 6.1

Model Fit Assessment

The fit of this hypothesized model to the data was evaluated using a standard set of fit indices.

Fit Index	Value / Interpretation

Chi-Square Test	$\chi^2(32) = 138.60$, $p < .001$. Rejects perfect fit, as expected with a large, non-normal sample.
GFI (Goodness of Fit Index)	0.975. Excellent fit (Value > 0.90 is considered good).
AGFI (Adjusted Goodness of Fit Index)	0.957. Excellent fit (Value > 0.85 is considered good).
SRMR (Standardized Root Mean Square Residual)	0.018. Excellent fit (Value < 0.10 is acceptable).

Table 3.3.2: CCA resultant fit indexes and acceptable values. Appendix 6.1

Conclusion on Model Fit

While the Chi-Square test was statistically significant, inflated by the data's non-normality as identified in Section 2.00, the other key fit indices strongly supported the model. The GFI and AGFI values well exceeded the thresholds for a good fit, and the SRMR indicated very small residuals. Therefore, the hypothesized 3-factor model was confirmed as an acceptable representation of the latent structure of student stress. After validating the underlying structure of the variables, the analysis shifted focus to using this structure to cluster the student population itself.

This part (3.3) was written by Clay Cleavinger

4. Student Clustering Analysis

While factor analysis revealed the underlying structure of stress, it did not identify how these constructs manifest within the student population. To bridge this gap from abstract structure to concrete typologies, a multi-method cluster analysis was employed. The objective was to move beyond variable-level insights and create data-driven student

profiles based on their comprehensive stress-related characteristics, thereby enabling the potential for targeted interventions.

4.1 K-Means Clustering

The first clustering method applied was K-Means using the standardized Euclidean distance between points. To properly identify the number of clusters, many different k's were estimated and evaluated using the Total Within-groups sum of squares and the elbow method. The graph for the elbow method displays a prominent elbow at k = 3.

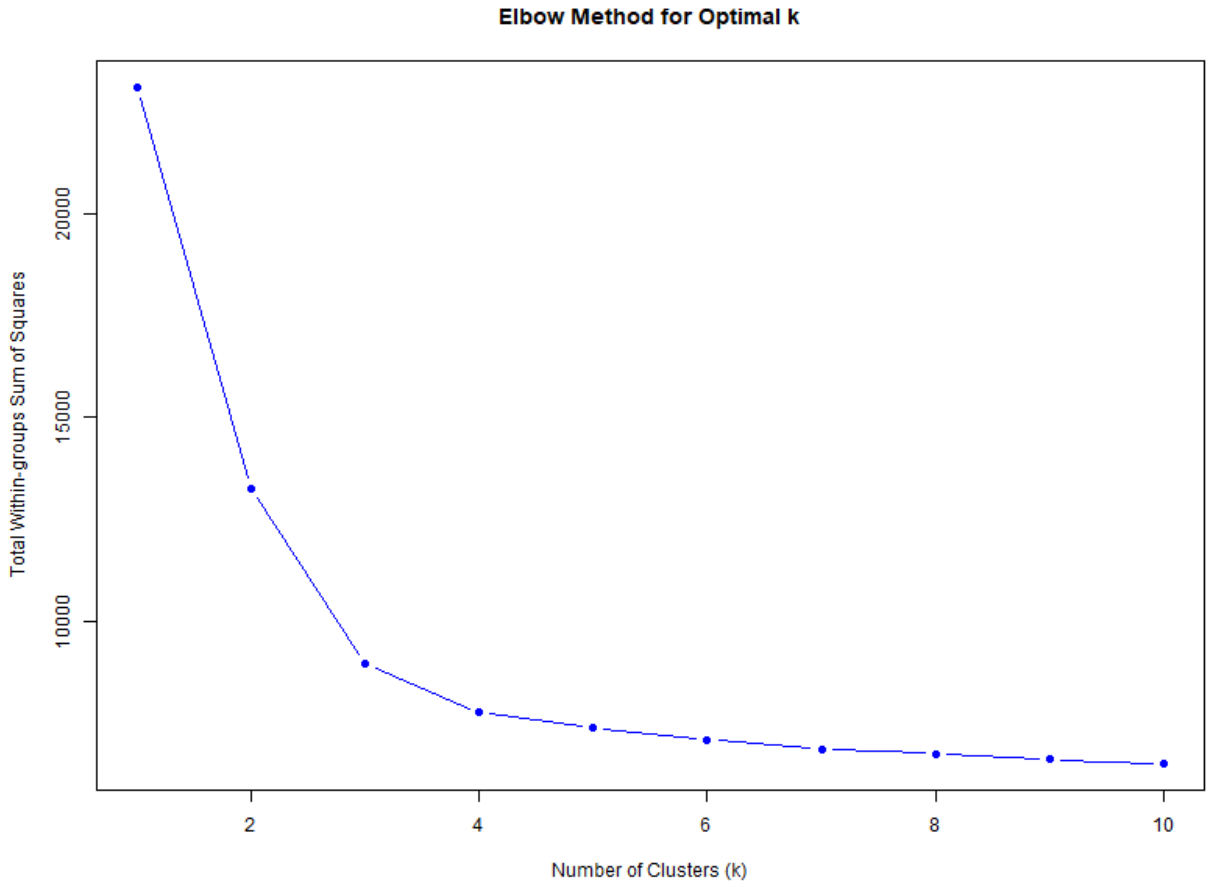


Figure 4.1.1: Graph of Elbow method used to find optimal k value for K-Means. Appendix 7.1

The resultant clusters from K-means present three well defined clusters spread along the PC1 Axis (General Stress). These clusters are mainly separated by the differences levels across clusters with a high, medium, and low stress level being apparent.

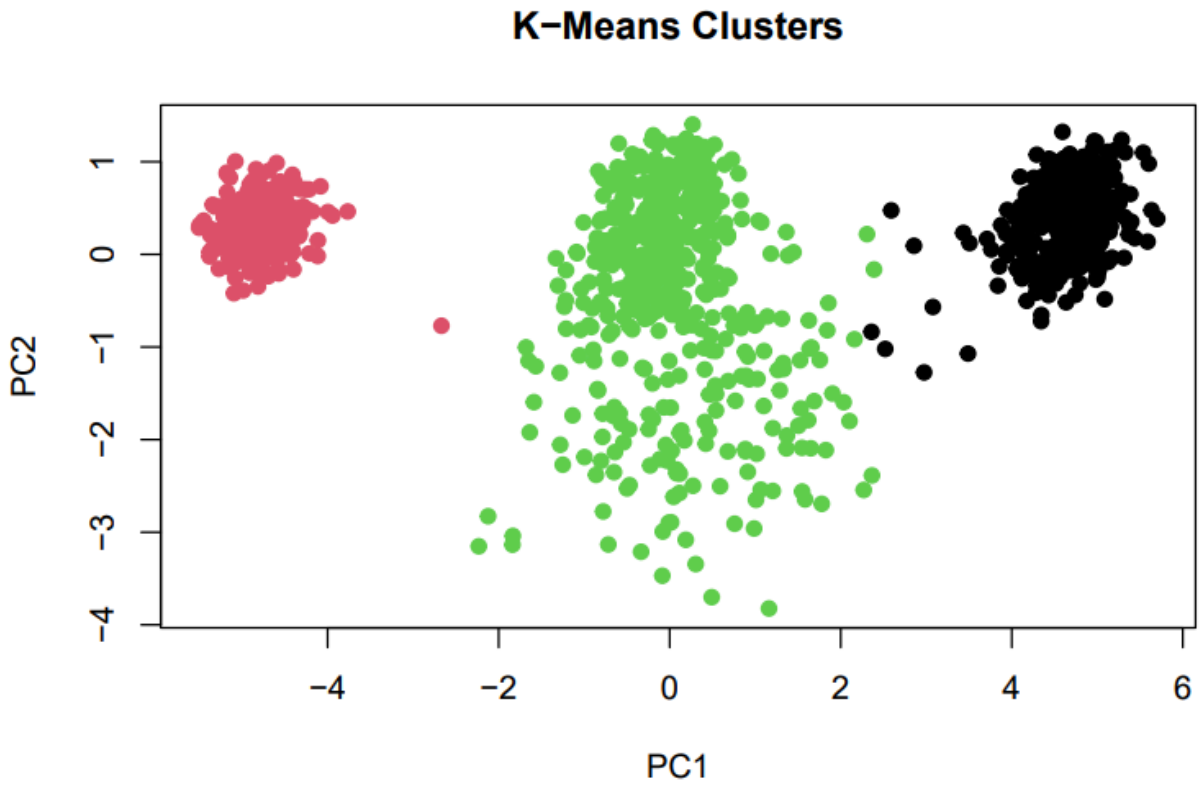


Figure 4.1.2: Graph of resultant clusters from K-Means. Cluster 1: Black, Cluster 2: Red, Cluster 3: Green. Appendix 7.1

4.2 Hierarchical Clustering

Hierarchical Clustering was used as a secondary technique to validate the clusters identified by K-Means. When the dendrogram is pruned to yield three clusters, the resulting groupings are comparable in size and closely aligned with K-Means which reinforces the robustness of this clustering solution. This similarity across clustering algorithms strengthens the confidence in the validity of the underlying cluster structure.

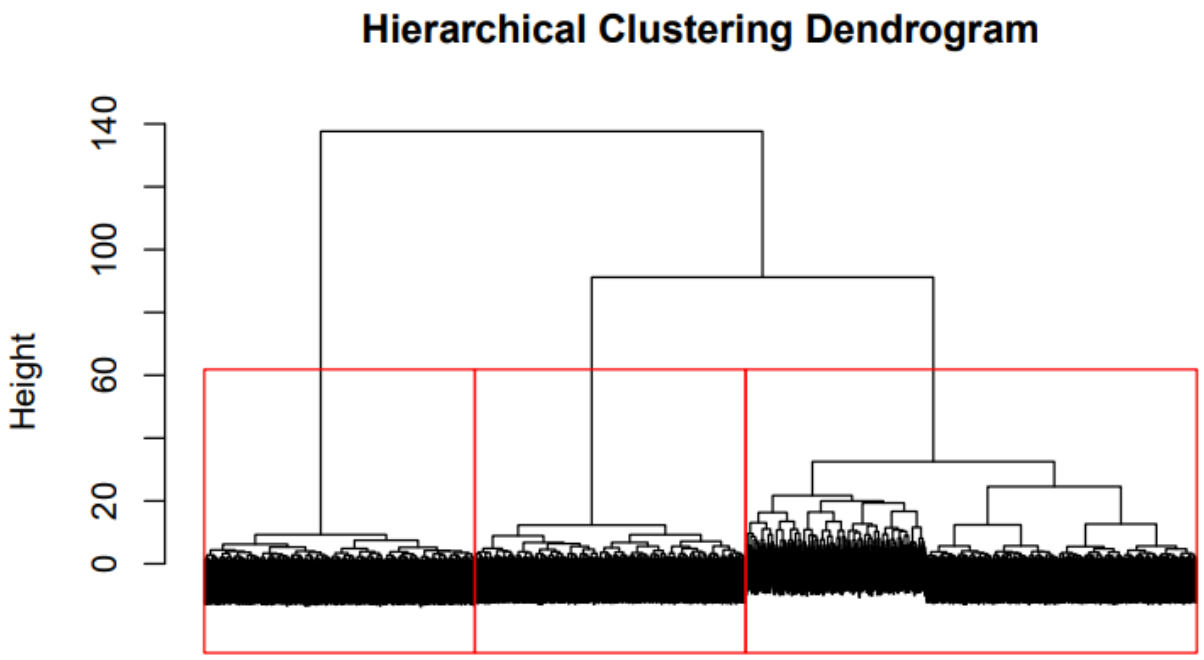


Figure 4.2.1: Hierarchical Dendrogram of students cut at Height = 60 and k=3 clusters.

Appendix 7.2

4.3 Model Based Clustering

Model Based clustering was also examined as a way to identify the underlying profiles of students. Unlike the previous methods, the model based identified 9 clusters as the optimal amount with the EVI model having the highest performance. In light of the data violating multivariate normality, it is highly probable that this method is insufficient to model our data in an interpretable way.

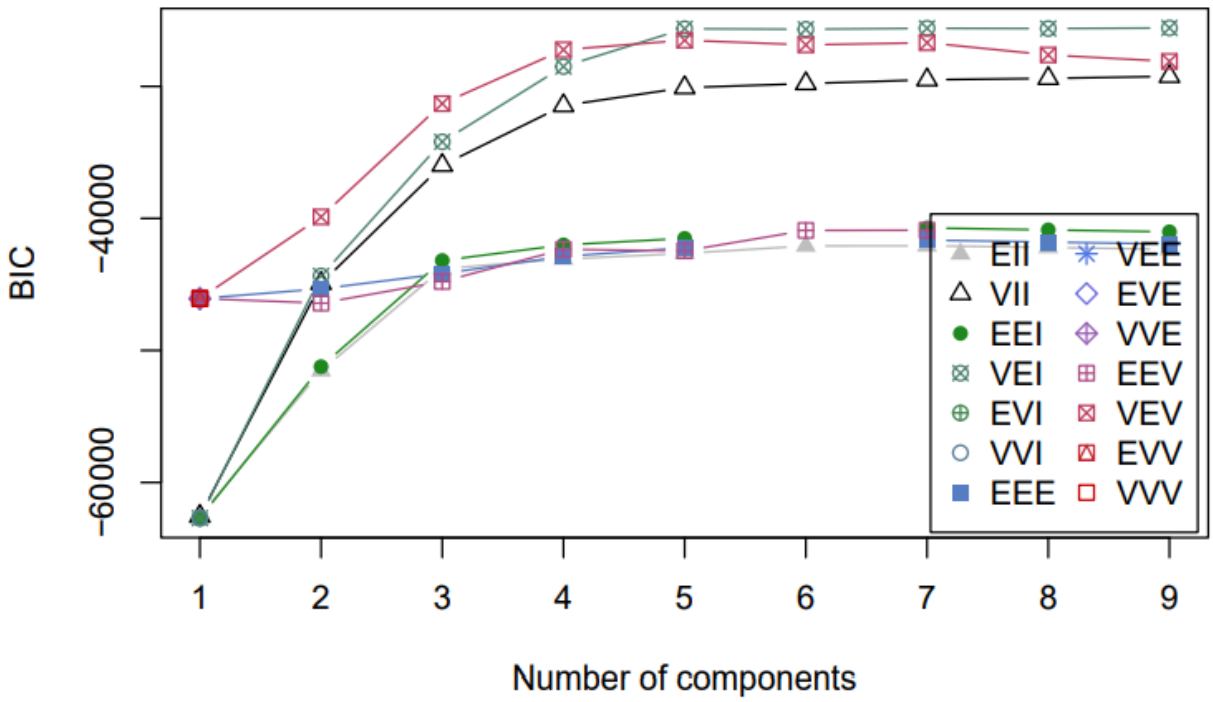


Figure 4.3.1: Model Based Clustering Results. Appendix 7.3

Comparison of Clustering Methods

Three distinct clustering algorithms were evaluated: K-Means (using the Elbow Method), Hierarchical Clustering (using Ward's method), and Model-Based Clustering (using the Bayesian Information Criterion, or BIC). While Model-Based clustering suggested a mathematically optimal 9-cluster solution, it is too granular for practical use in clustering the students. In contrast, both K-Means and Hierarchical methods consistently identified a 3-cluster solution. This solution was selected for final clustering as it offered a simpler, highly interpretable, and actionable clusters.

Final Student Cluster Profiles

The characteristics of the three final student clusters were analyzed by examining their average scores on key indicator variables.

• Cluster 1: High Risk. This group is defined by the highest average stress levels, the lowest academic performance (GPA), and the lowest self-esteem. They represent the most vulnerable segment of the student population.

• Cluster 2: Thriving. In stark contrast, this group exhibits the lowest average stress

levels, the highest GPA, and the highest self-esteem. This profile represents a resilient and well-supported population of students.
• Cluster 3: Moderate / At-Risk. This group falls in the middle, with moderate levels across stress, academic, and psychological metrics. These students may be experiencing situational stressors without yet falling into the high-risk category.

With distinct student segments identified, the next phase of the analysis focused on uncovering more specific relationships between different sets of variables.

This part (4) was written by Clay Cleavinger

5. Multidimensional Scaling (MDS) for Visual Validation

Multidimensional Scaling (MDS) was employed as a non-parametric visual confirmation of the parametric K-Means clustering results. The purpose of MDS is to create a two-dimensional "map" of the students where the distances between points correspond as closely as possible to their dissimilarities in the original high-dimensional data, providing an intuitive visual check of the data's inherent structure.

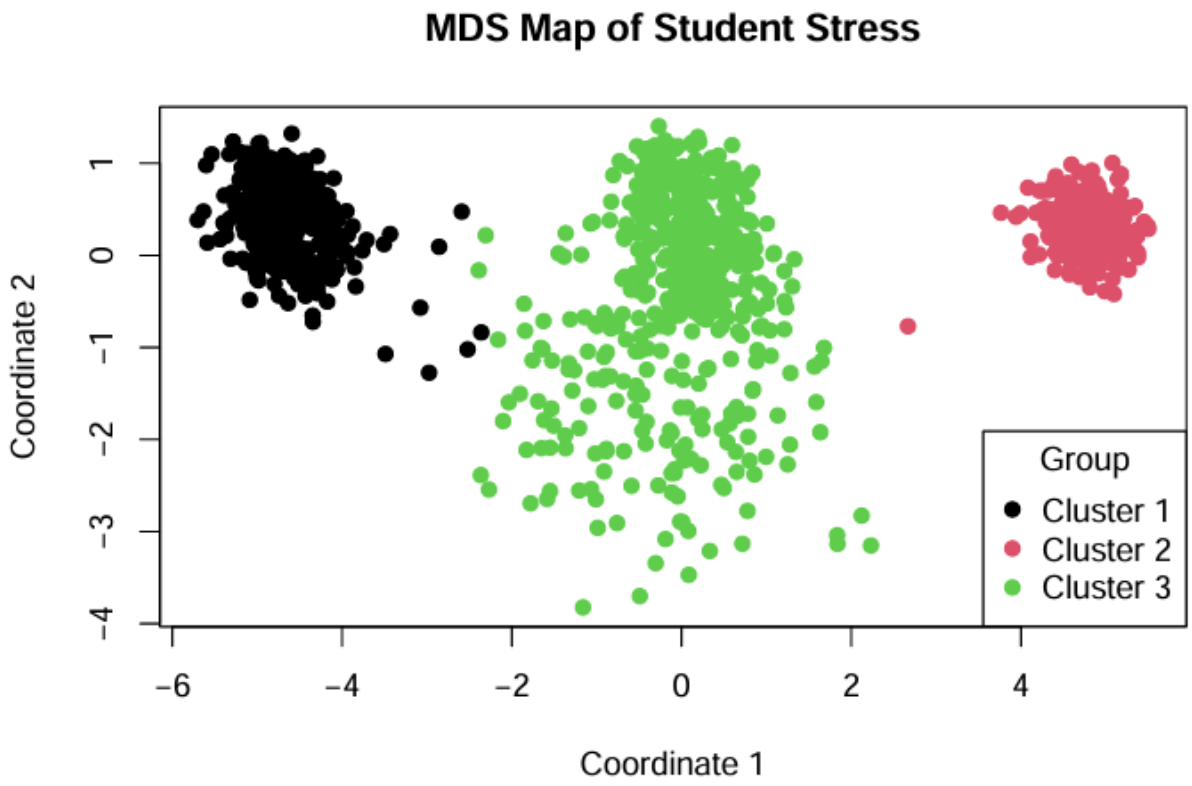


Figure 5.1.1: Mapping of the Multidimensional Scaling for the full students stress dataset. Appendix 8

MDS Results and Interpretation

The resulting MDS plot, when colored by K-Means cluster assignment, visually confirmed the distinct separation of the three student groups. This provides strong, independent evidence that the clusters are not artifacts of the K-Means algorithm but represent genuine, separable typologies within the student population.

Interpretation of MDS Axes

To give meaning to the map's layout, the two MDS coordinates were correlated with the original 20 variables. This analysis revealed the underlying meaning of the plot's axes:
• Coordinate 1: 'Overall Distress'. This axis showed a very high correlation with variables like stress_level (-0.90) and anxiety_level (-0.85) and a strong positive correlation with protective factors like self_esteem (0.83). This indicates when moving to the left on this map, stress increases.

• Coordinate 2: 'External/Social Factors'. This axis was most strongly correlated with variables such as social_support (0.51).

This interpretation confirms that the student clusters are meaningfully distinct across both internal psychological and external social dimensions, serving as a powerful independent validation of the core findings from the factor and cluster analyses.

This part (5) was written by Guus Bouwens.

6. Analysis of Associations Between Variables

After identifying the broad latent structures of stress and segmenting the student population, the analysis intended to uncover more specific, directed relationships between different types of stress-related factors. The following methods were used to move beyond general correlations and explore targeted associations between predefined sets of variables and categorical outcomes.

6.1 Canonical Correlation Analysis (CCA)

In our dataset, there are numerous sets of variables. These sets range from environmental, social, academic, and psychological factors. In order to test and see what set of variables strongly affect our target variable, stress level, we decided to choose one set of variables and run CCA to see the correlation with that set with the set of variables that stress levels fall in. For our CCA, “X” will represent our “psychological” variables. This set includes the following variables: anxiety level, self esteem, sleep quality, depression, and stress level which is our target. Our “Y” set of variables represents our “Environmental and Social Support” variables. The variables that are in his set are the following: social support, safety, basic needs, living conditions, and noise level. We will use these two sets of variables to see what associations we can make with them. Our CCA can be modeled by the heat maps that are presented below.

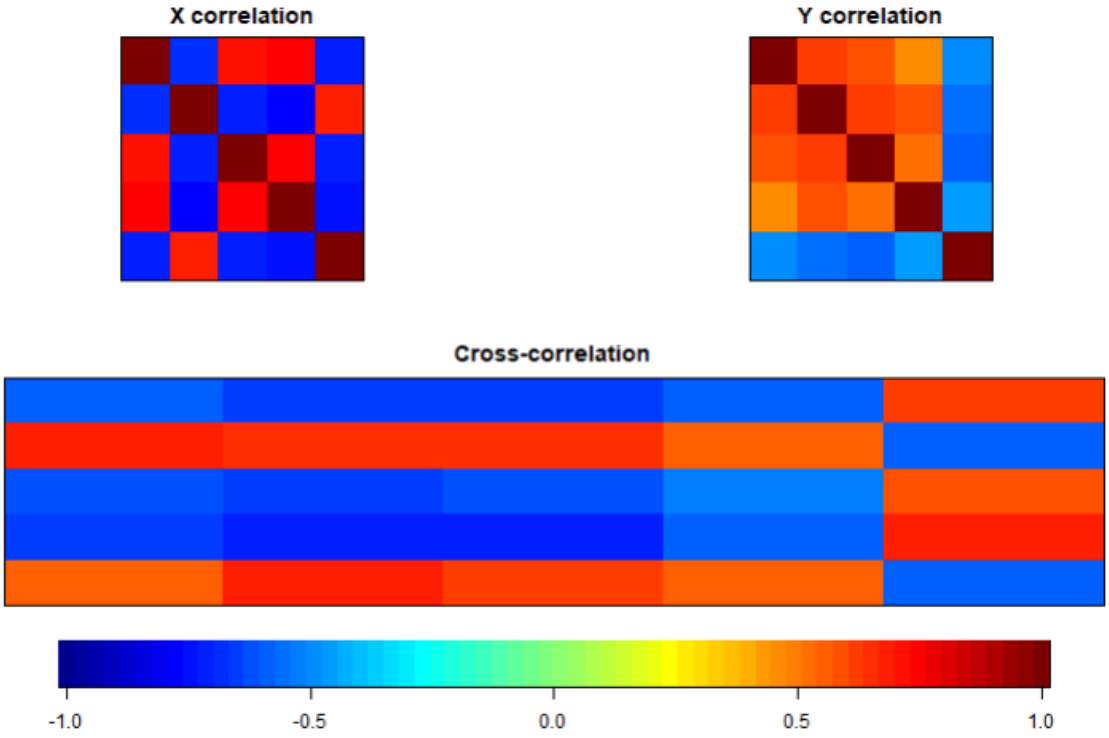


Diagram 6.1.1: CCA Heatmap. Appendix 6.2

The “X correlation” heatmap above represents the correlation between the variables in the psychological set of variables. The plot shows that there are very strong correlations that are both positive and negative in this set. The “Y correlation” heatmap represents the environmental and social support set of variables. The heatmap shows the same situation as the x correlation. There are strong correlations that are both positive and negative in this set of variables as well. Lastly, for the bottom heatmap, the “Cross-correlation” represents the CCA. Based on the plot, there is a strong relationship between the two sets of variables, illustrated by the dark blue and the dark red.

Moving on to the output of the CCA, the output was as follows:

0.884	0.273	0.079	0.069	0.012

Table 6.1.1: CCA Correlation Output

The output shows that there is a strong association between the two sets of variables, illustrated by 0.884. After that, the relationship falls off, likely due to noise. This shows that the correlation between the two sets of variables is dominated by a singular feature.

Moving on to the coefficients, the coefficients for the psychological set of variables is as follows:

Anxiety Level	-0.22
Self Esteem	0.24
Depression	-0.13
Stress Level	-0.4
Sleep Quality	.13

Table 6.1.2: Psychological Variable Set Coefficients. Appendix 6.2

This output for the psychological set of variables shows that high scoring observations will have high self esteem and sleep quality with low depression, anxiety, and stress. The high scores will represent those who are mentally stable.

Moving on to the next set of variables, the coefficients for the environmental and social support set of variables is as follows:

Social Support	0.23
Safety	0.29
Basic Needs	0.27
Living Conditions	0.18
Noise Level	-0.27

Table 6.1.3: Environment and Social Support Variable Set Coefficients. Appendix 6.2

For this set, a high scoring observation will be one that has great social support, all of their basic needs, a highly safe environment, and amazing living conditions with low noise levels. These high scores will represent those who have their basic safety and environmental needs.

To summarize all of the findings, the result of this CCA is that there is a strong association with those who are mentally stable with low stress with those who have great environments and social support. There seems to be a very strong linkage to what someone's environment is like and their social group to what their stress level will be.

This part, (6.1), was written by Tyler Campbell

6.2 Correspondence Analysis (CA)

Correspondence Analysis was employed to visualize the association between key categorical variables. To facilitate this, the continuous variables for stress_level and academic_performance were discretized into three categories each: Low, Medium, and High. The resulting CA plot positions categories in a two-dimensional space where proximity indicates a stronger association. The plot clearly showed that the "High Stress" category was located in close proximity to the "Low GPA" category, and vice-versa, providing a powerful visual confirmation of the negative association between these variables.

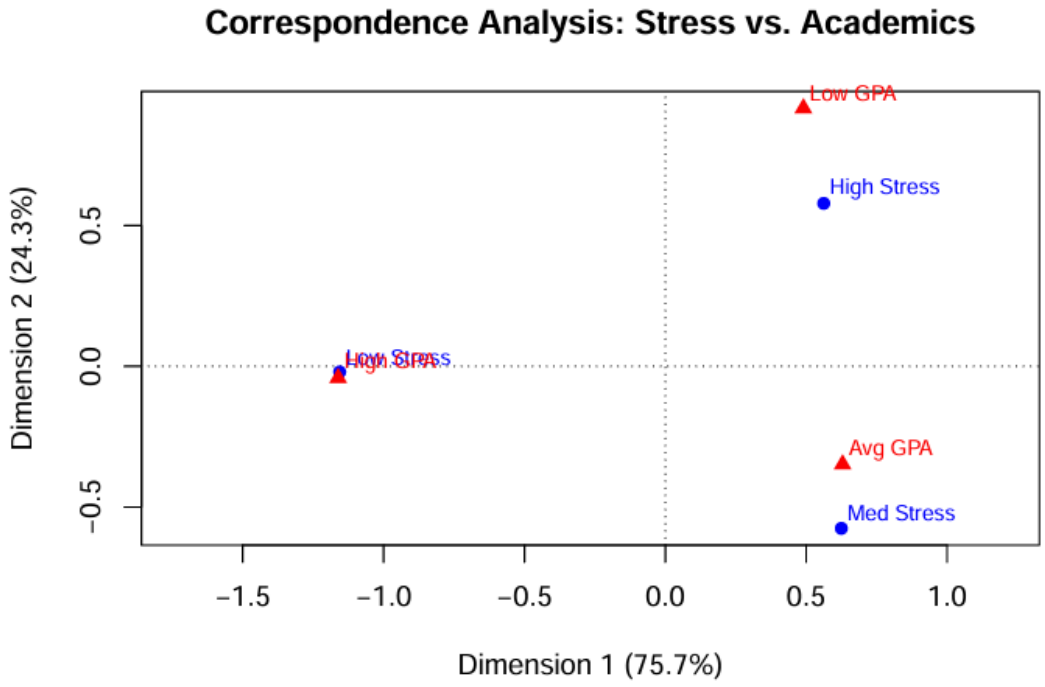


Figure 6.2.1: Visualization of the Correspondence Analysis of Stress Levels and Academic Performance on low, medium, and high divisions. Appendix 9

This part (6.2) was written by Guus Bouwens.

6.3 Association Rule Mining (Market Basket Analysis)

Market Basket Analysis was used to discover "if-then" association rules, identifying specific combinations of factors that strongly predict an outcome. The numerical data was first discretized into binary categories based on median splits. To move beyond tautological findings (e.g., {High_Anxiety} => {High_Stress}), rules containing direct psychological symptom indicators in the antecedent (left-hand side) were filtered post-generation. This focused the analysis on identifying combinations of non-obvious behavioral and environmental factors that predict high stress.

The most significant rule discovered was: {Low_Sleep, Low_SelfEsteem} => {High_Stress}. This finding was supported by two key metrics:

• Confidence: Students with this combination have a 78% probability of also

experiencing high stress.
• Lift: A student with this combination is 2.34 times more likely to have high stress than an average student.

The strategic implication is profound: it isolates a specific, actionable 'tipping point' for high stress that is more nuanced than obvious correlations, creating a target for preventative intervention.

The final section of this report synthesizes these multi-stage findings into a cohesive conclusion.

This part (6.3) was written by Guus Bouwens.

7. Conclusion

This report has detailed a multi-stage analytical journey designed to deconstruct the complex phenomenon of university student stress. The systematic application of complementary multivariate techniques provided a comprehensive and robust understanding that would be impossible to achieve through univariate approaches alone.

The key methodological findings of this analysis are summarized below:
• Validated Structure: The analysis successfully identified a theoretically sound 3-factor structure of student stress, comprising Psychological Well-being, Academic/External Pressure, and Environmental Support. This structure, first uncovered with EFA, was subsequently validated with CFA, which demonstrated an excellent model fit.
• Actionable Segmentation: Cluster analysis effectively partitioned the student population into three distinct profiles: "High Risk," "Moderate," and "Thriving", which was later validated by MDS. This segmentation provides a clear framework for allocating resources and designing targeted interventions.
• Critical Associations: Canonical Correlation Analysis found that students with a more stable psychological state (lower anxiety and stress) were strongly associated with a

more positive environment (higher safety and social support). Correspondence Analysis successfully confirmed the negative relation between higher stress and lower academic performance. Market Basket Analysis isolated a critical combination of risk factors. The discovery that Poor Sleep Quality combined with Low Self-Esteem makes a student over 2.3 times more likely to experience high stress provides a highly specific and actionable target for preventative initiatives.

Overall Implications

This report demonstrates how a systematic, multi-faceted analytical methodology can deconstruct a complex social issue like student stress into a solvable problem with clear, data-driven implications. By identifying the core components of stress, validating their structure, segmenting the student population, and isolating critical risk combinations, the analysis provides a roadmap for designing effective, targeted interventions. The findings show that a university's efforts could be most impactful not by applying a one-size-fits-all solution, but by focusing on specific, evidence-based needs, such as bolstering self-esteem and promoting healthy sleep hygiene for students who exhibit that high-risk profile.

Primary Methodological Limitation

The primary limitation encountered was the violation of the multivariate normality assumption. This directly impacted the Chi-square-based statistical tests in EFA and CFA, leading to statistically significant results that suggested poor model fit. However, the overall conclusions are considered robust because this limitation was mitigated by relying on descriptive fit indices, which are less sensitive to sample size and normality assumptions, and by the consistent convergence of findings across multiple, methodologically distinct techniques.

Future Directions

Future work should prioritize methodological robustness by employing techniques suited for multivariate non-normality, such as robust standard errors in CFA, to fully validate the adopted 3-factor structure. Research could also explore the statistically optimal but more complex 9-component segmentation suggested by Model-Based Clustering

(Mclust) to identify finer, more nuanced student risk profiles. Building on the successful findings of Rule #2, Market Basket Analysis (MBA) should be expanded to include academic and environmental variables (e.g., study load or social support) to uncover additional predictive "if-then" relationships. Lastly, a targeted analysis should investigate the excluded blood_pressure variable to determine if it holds meaningful correlations with stress factors when analyzed in a dedicated physiological context.

This part (7) was written by Guus Bouwens.

8. Appendix

Code

1. Load required libraries

```{r}
library(MVA)            # Multivariate analysis tools
library(HSAUR2)         # Data analysis tools
library(MASS)           # Methods including MDS
library(scatterplot3d)  # 3D plotting
library(sem)            # Structural Equation Modeling/CFA
library(semPlot)        # Path diagrams
library(corrplot)       # Correlation matrix visualization
library(dplyr)          # Data manipulation
library(CCA)            # Canonical Correlation Analysis
library(ca)             # Correspondence Analysis
library(mclust)         # Model-Based Clustering
library(arules)         # Market Basket Analysis
```

2. Data Cleaning and Visualization

2.1 Data Loading and Standardization

```{r}
# Load dataset
df <- read.csv("student_stress_dataset.csv")

# Remove 'blood_pressure' based on preliminary analysis
df <- df %>% dplyr::select(-blood_pressure)

# Check for missing values
cat("Missing values detected:", sum(is.na(df)), "\n")
```

```
# Standardize the data
df_scaled <- scale(df)
df_scaled <- as.data.frame(df_scaled)
```

2.2 Multivariate Normality Test

```{r}
x_center <- colMeans(df_scaled)
x_cov <- cov(df_scaled)
d2 <- mahalanobis(df_scaled, x_center, x_cov)

# Generate Q-Q Plot
quantiles <- qchisq((1:nrow(df_scaled) - 0.5) / nrow(df_scaled), df = ncol(df_scaled))
plot(quantiles, sort(d2),
     xlab = expression(paste(chi[20]^2, " Quantile")),
     ylab = "Ordered Squared Mahalanobis Distances",
     main = "Chi-Square Q-Q Plot")
abline(0, 1, col = "red")
```

2.3 Correlation Analysis

```{r}
cor_matrix <- cor(df_scaled)
corrplot(cor_matrix, method = "color", type = "upper",
         tl.col = "black", tl.cex = 0.6, order = "hclust",
         title = "Correlation Matrix of Student Stress Factors", mar=c(0,0,1,0))
```

3. Dimension Reduction: PCA

3.1 Performing PCA

```{r}
pca_result <- prcomp(df_scaled, scale. = TRUE)
pca_result$rotation <- -1 * pca_result$rotation
pca_result$x <- -1 * pca_result$x
summary(pca_result)
```

3.2 Determining the Number of Components

```{r}
# Calculate variance explained
var_explained <- pca_result$sdev^2 / sum(pca_result$sdev^2)

# Check Eigenvalues (Kaiser Criterion)
eigenvalues <- pca_result$sdev^2
print(round(eigenvalues, 2))
# Count how many are > 1
num_comp <- sum(eigenvalues > 1)
cat("Number of components with Eigenvalues > 1:", num_comp, "\n")

# Scree Plot
plot(var_explained, type = "b", pch = 19, col = "blue",
     xlab = "Principal Component",
     ylab = "Proportion of Variance Explained",
     main = "Scree Plot")
abline(h = 0.05, col="red", lty=2)
```

3.2 EFA

```{r}
# Fit 3-factor model with Varimax rotation
efa_fit <- factanal(df_scaled, factors = 3, rotation = "promax")

print(efa_fit)
```
```{r}

# Display loadings (suppressing small values for clarity)
prin

3.3 Interpreting the Components (Loadings)
```{r}



# Extract loadings for the first 3 PCs
loadings <- pca_result$rotation[, 1:3]
print(round(loadings, 2))
```

3.4 Biplot Visualization

```{r}
# Create a vector of dots instead of row numbers
biplot(pca_result, scale = 0, cex = 0.6,
       xlabs = rep("·", nrow(df_scaled)), # Replaces numbers with small dots
       col = c("grey", "red"),
       main = "PCA Biplot (PC1 vs PC2)")
```
3.5 3D Visualization
```{r}

```
scores <- pca_result$x[, 1:3]
s3d <- scatterplot3d(scores, color = "blue", pch = 19,
                     main = "3D Plot of First 3 PCs",
                     angle = 45, grid=TRUE, box=FALSE)
```

4. Canonical Correlation Analysis (CCA)

```{r}
# Set X: Psychological Variables
X <- df_scaled[, c("anxiety_level", "self_esteem", "depression", "stress_level",
"sleep_quality")]

# Set Y: Environmental & Social Support Variables
Y <- df_scaled[, c("social_support", "safety", "basic_needs", "living_conditions",
"noise_level")]

# Perform CCA
cca_result <- cc(X, Y)

# Canonical correlations
print(cca_result$cor)
print(cca_result$xcoef)
print(cca_result$ycoef)
```

5. Exploratory Factor Analysis (EFA)
```{r}
# Fit 3-factor model with Varimax rotation
efa_fit <- factanal(df_scaled, factors = 3, rotation = "promax")
```

```
print(efa_fit)
```

```{r}

# Display loadings (suppressing small values for clarity)
print(efa_fit$loadings, cutoff = 0.45)
```

6.1 Confirmatory Factor Analysis (CFA)

```{r}
# Define model syntax
cfa_model <- specifyModel(text = "
  Psychological -> anxiety_level, lam1, NA
  Psychological -> depression, lam2, NA
  Psychological -> self_esteem, lam3, NA
  Psychological -> stress_level, lam4, NA

  Pressure      -> peer_pressure, lam5, NA
  Pressure      -> study_load, lam6, NA
  Pressure      -> future_career_concerns, lam7, NA

  Support       -> social_support, lam8, NA
  Support       -> safety, lam9, NA
  Support       -> basic_needs, lam10, NA

  Psychological <-> Psychological, NA, 1
  Pressure <-> Pressure, NA, 1
  Support <-> Support, NA, 1
```

```
  Psychological <-> Pressure, rho1, NA
  Psychological <-> Support, rho2, NA
  Pressure <-> Support, rho3, NA

  anxiety_level <-> anxiety_level, the1, NA
  depression <-> depression, the2, NA
  self_esteem <-> self_esteem, the3, NA
  stress_level <-> stress_level, the4, NA
  peer_pressure <-> peer_pressure, the5, NA
  study_load <-> study_load, the6, NA
  future_career_concerns <-> future_career_concerns, the7, NA
  social_support <-> social_support, the8, NA
  safety <-> safety, the9, NA
  basic_needs <-> basic_needs, the10, NA
")

# Fit SEM using the correlation matrix
C <- cor(df_scaled)
fit <- sem(cfa_model, C, N = nrow(df_scaled))

# Display Summary
summary(fit)

# Visualize Path Diagram
semPaths(fit, rotation = 2, "est",
         main="CFA Path Diagram",
         whatLabels = "est",
         node.width = 0.8,
         edge.label.cex = 0.8)
```

6.2 Canonical Correlation Analysis

```{r}
# Set X: Psychological Variables
X <- df_scaled[, c("anxiety_level", "self_esteem", "depression", "stress_level",
"sleep_quality")]

# Set Y: Environmental & Social Support Variables
Y <- df_scaled[, c("social_support", "safety", "basic_needs", "living_conditions",
"noise_level")]

# Perform CCA
cca_result <- cc(X, Y)

# Canonical correlations
print(cca_result$cor)
print(cca_result$xcoef)
print(cca_result$ycoef)
```

```{r}
#Assessing Model Fit

# IMPORTANT: We must tell the 'sem' package which indices to calculate
options(fit.indices = c("GFI", "AGFI", "SRMR"))

# Extract fit indices from summary
summ <- summary(fit)

# Calculate the p-value manually (1 - ChiSquare Distribution)
# (The 'chisq' value in the summary object is the Statistic, not the p-value)
p_value <- 1 - pchisq(summ$chisq, summ$df)
```

```
# Print results clearly
cat("--- Model Fit Assessment ---\n")
cat("Chi-Square Statistic:", round(summ$chisq, 3), "\n")
cat("Degrees of Freedom:", summ$df, "\n")
cat("Chi-Square p-value:", format.pval(p_value, digits=4), "\n")
cat("----------------------------\n")
cat("GFI (Goodness of Fit):", round(summ$GFI, 3), "\n")
cat("AGFI (Adjusted GFI):", round(summ$AGFI, 3), "\n")
cat("SRMR (Std. Root Mean Square Residual):", round(summ$SRMR, 3), "\n")
```

7. Cluster Analysis

7.1 K-Means Clustering (The Elbow Method)

```{r}
wgss <- numeric(10)
for (i in 1:10) {
  km <- kmeans(df_scaled, centers = i, nstart = 20)
  wgss[i] <- km$tot.withinss
}

plot(1:10, wgss, type = "b", pch = 19, col = "blue",
     xlab = "Number of Clusters",
     ylab = "Within groups sum of squares",
     main = "Elbow Method for Optimal k")
```

```{r}
set.seed(123)
```

```
km_fit <- kmeans(df_scaled, centers = 3, nstart = 25)
# Visualize Clusters on PCA
plot(pca_result$x[,1], pca_result$x[,2], col = km_fit$cluster, pch = 19,
     xlab = "PC1", ylab = "PC2", main = "K-Means Clusters")
```

7.2 Hierarchical Clustering

```{r}
d_matrix <- dist(df_scaled)
hc_fit <- hclust(d_matrix, method = "ward.D2")

# Plot Dendrogram
plot(hc_fit, labels = FALSE, main = "Hierarchical Clustering Dendrogram", xlab = "", sub
= "")
rect.hclust(hc_fit, k = 3, border = "red")

# Cut tree into 3 clusters
hc_clusters <- cutree(hc_fit, k = 3)
table(hc_clusters)
```

7.3 Model-Based Clustering (Mclust)

```{r}
# Fit Model-Based Clustering
mc_fit <- Mclust(df_scaled)

# Summary of the best model found
summary(mc_fit)
```

```
# Plot BIC to visualize model selection
plot(mc_fit, what = "BIC")

# Plot classification
# plot(mc_fit, what = "classification") # Commented out to save space in PDF
```

7.4 Overall Cluster Profiles:

```{r}
# Validate cluster characteristics
df_scaled %>%
  mutate(Cluster = km_fit$cluster) %>%
  group_by(Cluster) %>%
  summarise(
    Avg_Stress = mean(stress_level),
    Avg_GPA = mean(academic_performance),
    Avg_SelfEsteem = mean(self_esteem)
  ) %>%
  print()
```

8. Multidimensional Scaling (MDS)

```{r}
# 1. Calculate Euclidean Distance Matrix
d <- dist(df_scaled)

# 2. Perform Classical MDS
mds_fit <- cmdscale(d, k = 2, eig = TRUE)
```

```
# Correlate MDS coordinates with original variables to interpret axes
cor(df_scaled, mds_fit$points) %>% round(2)

# 3. Visualize
plot(mds_fit$points[,1], mds_fit$points[,2],
     col = km_fit$cluster,
     pch = 19,
     xlab = "Coordinate 1", ylab = "Coordinate 2",
     main = "MDS Map of Student Stress")
legend("bottomright", legend = paste("Cluster", 1:3),
       col = 1:3, pch = 19, title="Group")
```

9. Correspondence Analysis (CA)

```{r}
# Discretize variables
df$Stress_Cat <- cut(df$stress_level, breaks = 3, labels = c("Low Stress", "Med Stress",
"High Stress"))
df$Academic_Cat <- cut(df$academic_performance, breaks = 3, labels = c("Low GPA",
"Avg GPA", "High GPA"))

# Create contingency table
cont_table <- table(df$Stress_Cat, df$Academic_Cat)
print(cont_table)

# Perform CA
ca_result <- ca(cont_table)
plot(ca_result, main = "Correspondence Analysis: Stress vs. Academics")
```

10. Market Basket Analysis (Association Rules)

```{r}
# 1. Discretize using Medians (more robust to outliers than Mean)
df_mba <- df %>%
  transmute(
    High_Anxiety = ifelse(anxiety_level > median(anxiety_level), 1, 0),
    High_Stress = ifelse(stress_level > median(stress_level), 1, 0),
    Low_Sleep = ifelse(sleep_quality < median(sleep_quality), 1, 0),
    Low_SelfEsteem = ifelse(self_esteem < median(self_esteem), 1, 0),
    High_Depression = ifelse(depression > median(depression), 1, 0),
    # For bullying, keep > 0 as it is likely a count or binary
    Bullying_History = ifelse(bullying > 0, 1, 0)
  )

# 2. Convert to transactions matrix
trans_matrix <- as.matrix(df_mba)
trans <- as(trans_matrix, "transactions")

# 3. Generate Rules
rules <- apriori(trans, parameter = list(supp = 0.1, conf = 0.5, minlen = 2))

# 4. Filter for specific interesting rules
# Find what causes High Stress, but remove obvious tautologies like 'High_Anxiety' or
'High_Depression'
interesting_rules <- subset(rules,
                            subset = rhs %in% "High_Stress" &
                            !(lhs %in% "High_Anxiety") &
                            !(lhs %in% "High_Depression"))

# Sort by lift and inspect
```

```
inspect(head(sort(interesting_rules, by = "lift"), 5))
```

Works Cited

Andersen, R., Holm, A., & Côté, J. E. (2021). The Student Mental Health Crisis: Assessing Psychiatric and Developmental Explanatory Models. *Journal of Adolescence*, *86*(1), 101–114. https://doi.org/10.1016/j.adolescence.2020.12.004

Spagert, L., Janssen, C., & Geigl, C. (2022). Mental health indicators and their lifestyle associations in German students: A gender-specific multivariable analysis. *BMC Public Health*, *22*(1). https://doi.org/10.1186/s12889-022-13777-7