7/12/2023

Unravelling the Complexities of Student Stress Factors

ISQS 5346 – Final Project Report

Team 2 Members:

Arshdeep Kaur, Karan Bhosale, Nahid Ferdous, Priyanka Chahande

Table of Contents

PROJECT DESCRIPTION

(This section is authored by Arshdeep Kaur and Karan Bhosale)

Problem Statement

Despite the essential role education plays in shaping futures, student stress has become a pressing concern. Our problem statement is to explore, analyze, and comprehend the impact of psychological, physiological, social, environmental, and academic factors on student well-being to guide wellness approaches and treatments that encourage a more supportive and healthy learning environment. The need is to unlock meaningful insights from our dataset, providing actionable information to help alleviate stress and enhance the overall student experience.

Dataset Overview

The data was sourced from Kaggle.com [1]. Our dataset includes about 20 variables chosen for their scientific value, provide a thorough understanding of the intricacies of student life. In addition, it provides a systematic approach to examine and understand the varied experiences that students have during their academic careers. Like in the table below we have 20 manifest variables (independent variables) categorized under 5 latent variables. The dependent variable is the stress_level for which the analysis is done.

Psychological Factors:	*anxiety_level*	*self_esteem*	*mental_health_history*	*depression*
Physiological Factors:	*headache*	*blood_pressure*	*sleep_quality*	*breathing_problem*
Environmental Factors:	*noise_level*	*living_conditions*	*safety*	*basic_needs*
Academic Factors:	*academic_performance*	*study_load*	*teacher_student_relationship*	*future_career_concerns*
Social Factor:	*social_support*	*peer_pressure*	*extracurricular_activities*	*bullying*

Table 1. Manifest Variables and their Latent Factors

In each observation the anxiety level ranges from 0 to 21 and comes from the student's score filling the GAD-7 questionnaire. The Self-esteem variable has values ranging from 0 to 30 and comes from the Rosenberg Self Esteem Scale questionnaire. Mental Health History variable is 0 if student has no mental health history and 1 if student had mental health history. The Depression variable has values from 0 to 27 and is the student's score on the Patient Health Questionnaire (PHQ-9). Other features mostly range from 0 to 5 considering 0, 1 to be low, 2, 3 to be mid, and 4, 5 to be high [2].

DATA CLEANING AND VISUALIZATION

(This section is authored by Nahid Ferdous)

Errors, missing numbers, inconsistencies, and outliers may be present in raw data that has been gathered from THE SOURCE. To guarantee that the data is accurate and trustworthy, data cleaning entails locating and fixing these problems.

We can spot patterns, trends, or anomalies more rapidly when you visualize the data. Visualization tools frequently facilitate the identification of mistakes or anomalies in the data.

```
Rows: 1,100
Columns: 22
$ student_id                  <chr> "S1", "S2", "S3", "S4", "S5", "S6", "S7", "S8", "S9", "S10", "S11", "S12", "S13", …
$ anxiety_level               <int> 14, 15, 12, 16, 16, 20, 4, 17, 13, 6, 17, 17, 5, 9, 2, 11, 6, 7, 11, 21, 3, 18, 7,…
$ self_esteem                 <int> 20, 8, 18, 12, 28, 13, 26, 3, 22, 8, 12, 15, 28, 23, 28, 21, 28, 25, 23, 1, 27, 1,…
$ mental_health_history       <int> 0, 1, 1, 1, 0, 1, 0, 1, 1, 0, 1, 1, 0, 1, 0, 0, 0, 0, 0, 1, 0, 1, 0, 1, 1, 0, 1, 0…
$ depression                  <int> 11, 15, 14, 15, 7, 21, 6, 22, 12, 27, 25, 22, 8, 24, 3, 14, 1, 3, 12, 25, 0, 21, 5…
$ headache                    <int> 2, 5, 2, 4, 2, 3, 1, 4, 3, 4, 4, 3, 1, 4, 1, 3, 1, 1, 3, 4, 1, 4, 1, 3, 3, 1, 5, 2…
$ blood_pressure              <int> 1, 3, 1, 3, 3, 3, 2, 3, 1, 3, 3, 3, 2, 3, 2, 1, 2, 2, 1, 3, 2, 3, 2, 3, 1, 2, 3, 3…
$ sleep_quality               <int> 2, 1, 2, 1, 5, 1, 4, 1, 2, 1, 1, 1, 4, 1, 4, NA, 4, 4, 2, 1, 4, 1, 4, 1, 2, 5, 1, …
$ breathing_problem           <int> 4, 4, 2, 3, 1, 4, 1, 5, 4, 2, 3, 5, 2, 0, 2, 4, 2, 2, 2, 4, 1, 3, 1, 4, 2, 2, 5, 1…
$ noise_level                 <int> 2, 3, 2, 4, 3, 3, 1, 3, 3, 0, 4, 5, 2, 1, 1, 2, 1, 2, 3, 4, 1, 5, 1, 4, 3, 2, 3, 5…
$ living_conditions           <int> 3, 1, 2, 2, 2, 2, 4, 1, 3, 5, 2, 2, 3, 2, 3, 2, 4, 4, 2, 1, 3, 1, 3, 2, 2, 4, 2, 1…
$ safety                      <int> 3, 2, 3, 2, 4, 2, 4, 1, 3, 2, 1, 1, 5, 4, 4, 2, 5, 5, 3, 2, 5, 1, 5, 1, 2, 5, 2, 2…
$ basic_needs                 <int> 2, 2, 2, 2, 3, 1, 4, 1, 3, 2, 1, 1, 5, 3, 4, 2, 4, 4, 3, 1, 4, 2, 5, 2, 3, 4, 1, 5…
$ academic_performance        <int> 3, 1, 2, 2, 4, 2, 5, 1, 3, 2, 1, 1, 5, 1, 4, 3, 5, 4, 2, 1, 5, 2, 4, 1, 3, 4, 1, 4…
$ study_load                  <int> 2, 4, 3, 4, 3, 5, 1, 3, 3, 2, 3, 3, 2, 2, 2, 3, 1, 2, 3, 5, 2, 5, 2, 3, 2, 1, 3, 0…
$ teacher_student_relationship <int> 3, 1, 3, 1, 1, 2, 4, 2, 2, 1, 1, 1, 4, 3, 5, 3, 5, 5, 2, 2, 5, 1, 5, 1, 3, 4, 1, 2…
$ future_career_concerns      <int> 3, 5, 2, 4, 2, 5, 1, 4, 3, 5, 4, 4, 1, 3, 1, 3, 1, 1, 2, 5, 1, 4, 1, 4, 2, 1, 4, 2…
$ social_support              <int> 2, 1, 2, 1, 1, 1, 3, 1, 3, 1, 1, 1, 3, 0, 3, 2, 3, 3, 3, 1, 3, 1, 3, 1, 3, 3, 1, 1…
$ peer_pressure               <int> 3, 4, 3, 4, 5, 4, 2, 4, 3, 5, 4, 5, 1, 1, 1, 3, 2, 1, 3, 4, 1, 4, 1, 5, 3, 2, 5, 1…
$ extracurricular_activities  <int> 3, 5, 2, 4, 0, 4, 2, 4, 2, 3, 4, 5, 1, 0, 2, 2, 2, 1, 2, 4, 2, 4, 2, 4, 3, 1, 5, 2…
$ bullying                    <int> 2, 5, 2, 5, 5, 5, 1, 5, 2, 4, 5, 4, 1, 1, 1, 2, 1, 1, 3, 5, 1, 5, 1, 4, 2, 1, 4, 2…
$ stress_level                <int> 1, 2, 1, 2, 1, 2, 0, 2, 1, 1, 2, 2, 0, 2, 0, 1, 0, 0, 1, 2, 0, 2, 0, 2, 1, 0, 2, 0…
```

Fig.1: Insight into Data Type of Variables

For our dataset, the above visualization (fig:01) provides insights into the data types of the variables, their range, and the dimensions of the dataset. We have a dataset comprising 1100 observations and 22 variables. Among these, only one variable is of character (chr) type, while the remaining 21 variables are of integer (int) type.

In the dataset we did have missing values in the variables sleep_quality, headache, depression, basic_needs and living_conditions as shown in the figure below.

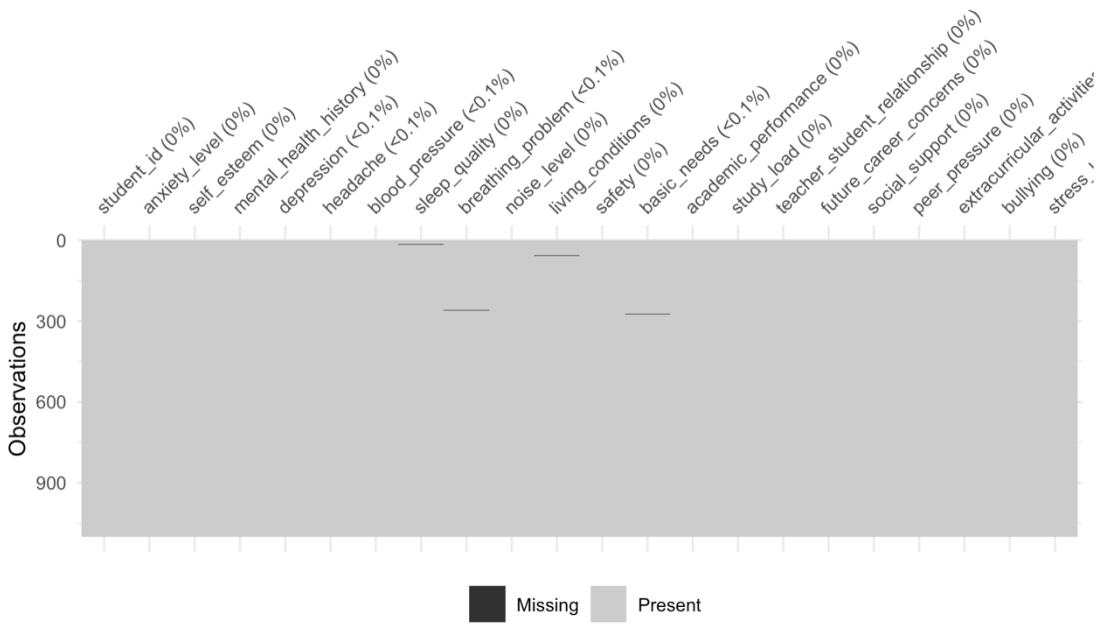


Fig.2: Illustrate Missing Values

Here (Fig:2) we illustrate the missing values, with black markers indicating their presence.

student_id	anxiety_level	self_esteem	mental_health_history
0	0	0	0
depression	headache	blood_pressure	sleep_quality
1	1	1	2
breathing_problem	noise_level	living_conditions	safety
1	0	3	0
basic_needs	academic_performance	study_load	teacher_student_relationship
1	0	0	0
future_career_concerns	social_support	peer_pressure	extracurricular_activities
0	0	0	0
bullying	stress_level		
0	0		

Fig.3: Data Used for Analysis

From the analysis, it's evident that the variables Depression, Breathing Problem, Basic Needs, Headache, Blood Pressure, Living Conditions, and Sleep Quality contain a small number of missing values. Considering our dataset comprises 1100 observations, we prefer not to discard any data points. Therefore, we will proceed with column-based mean imputation to address these missing values effectively.

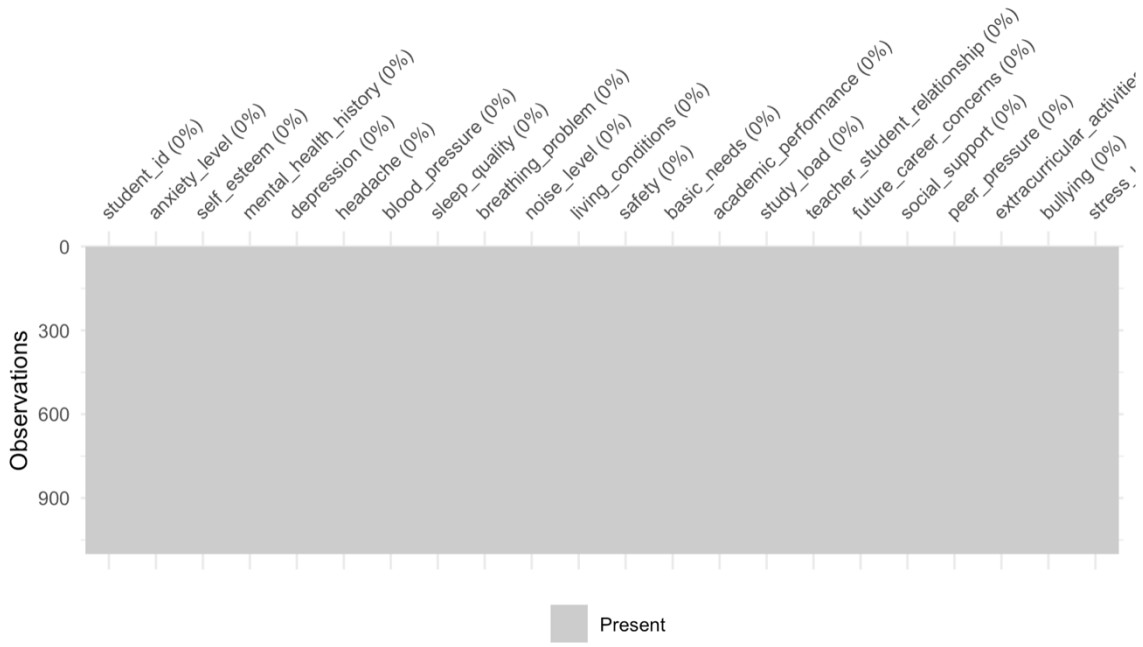


Fig.4: Mean Imputed Data

Following mean imputation, we reviewed our data and confirmed that there are no longer any missing values. (Fig: 4)

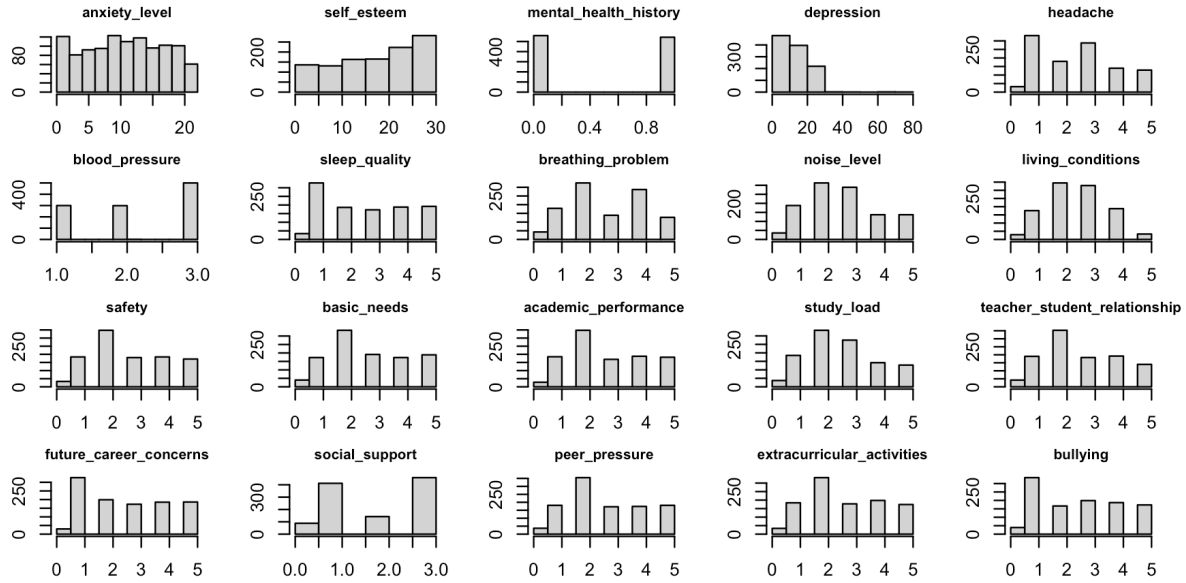


Fig.5: Histogram Visualization: Variable Distribution

This histogram visualization (Fig: 5) helps us analyze the distribution of our variables. It

appears that most variables do not follow a normal distribution, except for the 'anxiety_level'. Additionally, the visualization provides insights into outliers; notably, it is evident that the 'depression level' variable contains some outliers.

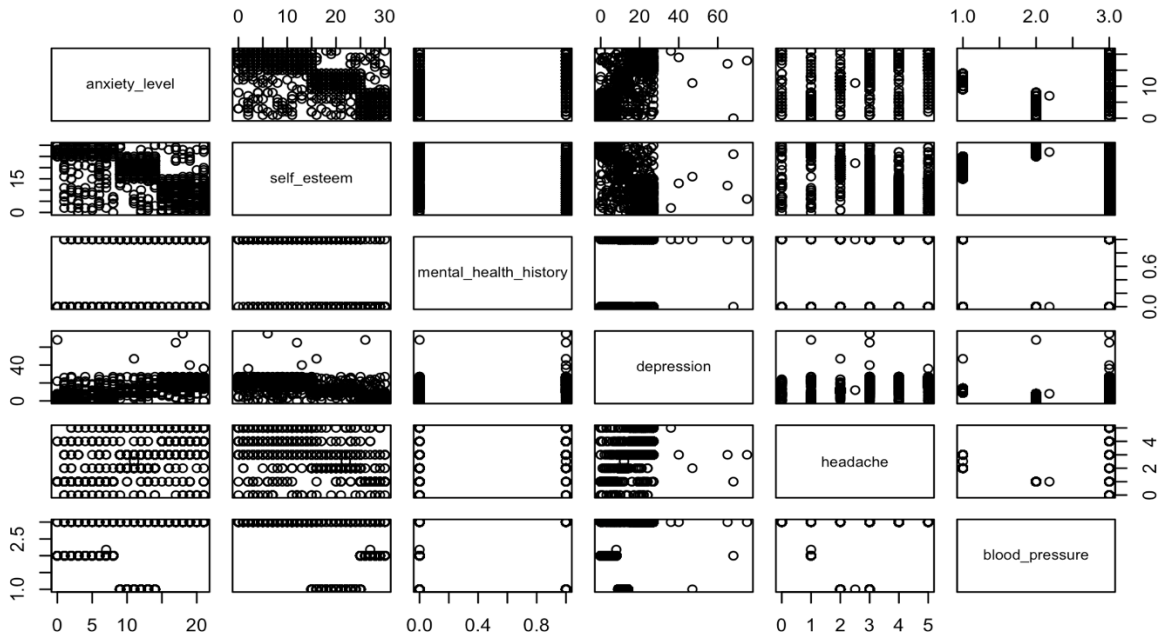


Fig.6: Outlier Detected Visualization

From this visualization (Fig: 6) we can say that "depression" variable has some outlines.

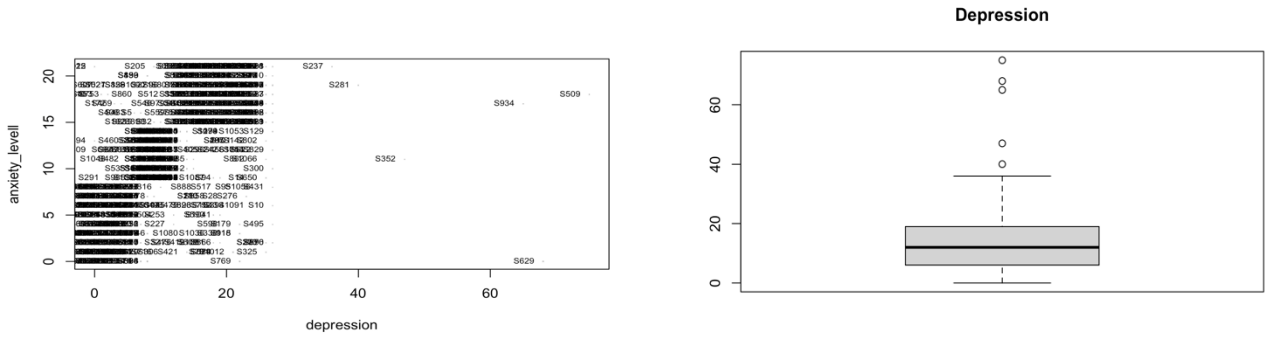


Fig.7: Specified Outlier Visualization

Here Depression variable contain "S281","S352", "S629","S934","S509", "S237" outliers. We got this using this visualization. (Fig:7)

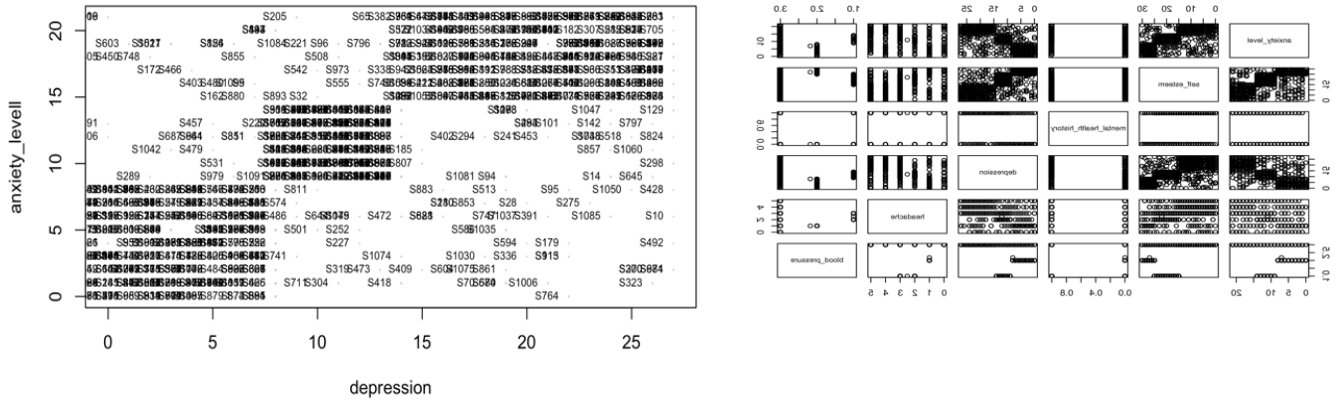


Fig.8: Specified Outlier Visualization

We can now ascertain that the variable representing depression does not exhibit any clear outliers.

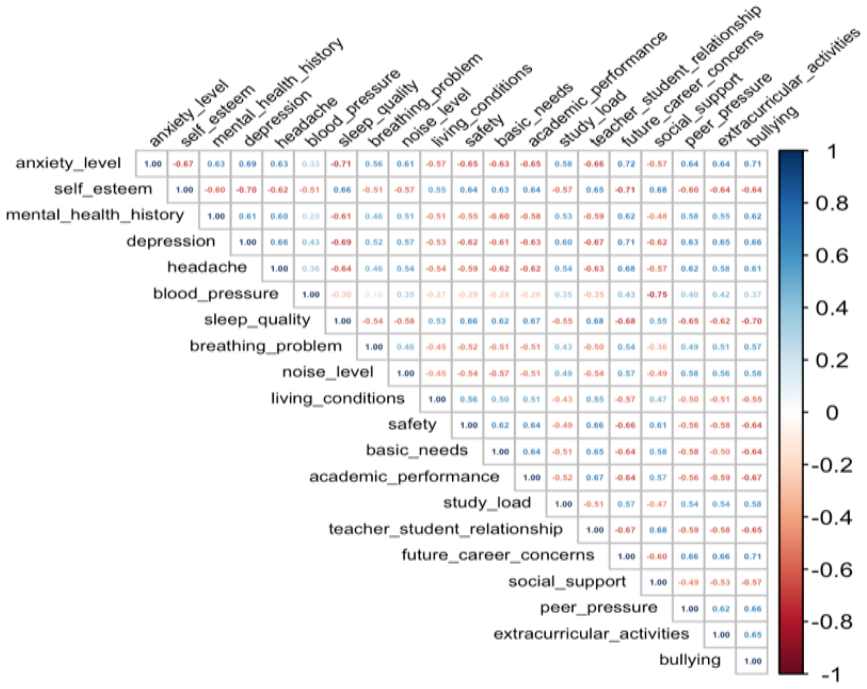


Fig.9: Correlation Matrix

Based on the correlation matrix (Fig: 9), it's evident that the variables in our dataset are highly correlated with each other. Given the large number of variables, all of which are numerical except for one, this dataset appears to be well-suited for multivariate analysis. This analysis will allow us to explore the complex interrelationships between these variables in depth.

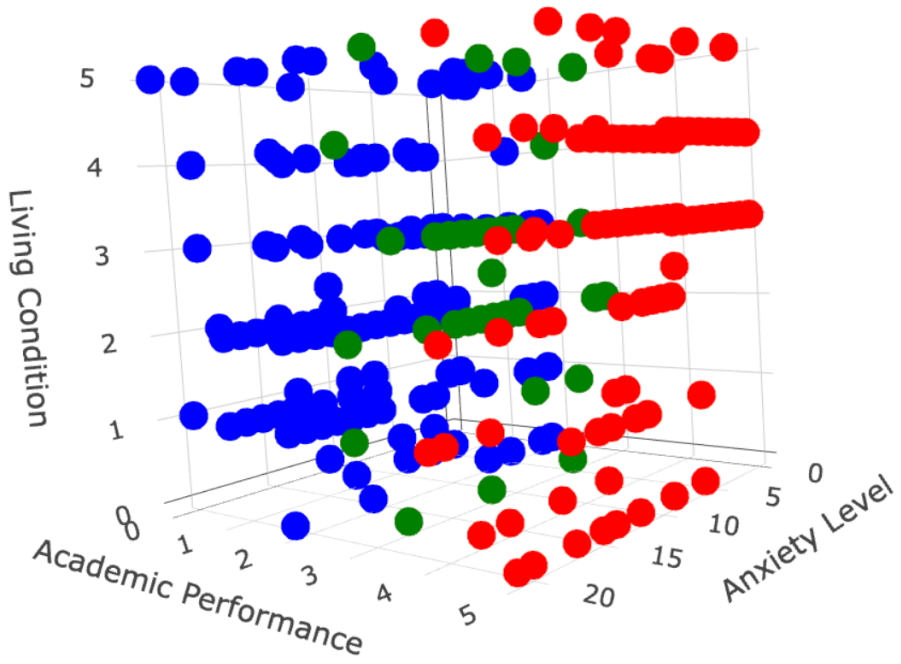


Fig.10: 3-D Description of Variables

In this 3D visualization, the blue points represent low academic performance, the green points represent medium academic performance, and the red points indicate high academic performance. This visualization suggests that if a student exhibits a low anxiety level and their living condition is between mild to high, then there's a higher likelihood of the student achieving a high academic score.

DIMENSION REDUCTION ANALYSIS

(This section is authored by Nahid Ferdous)

The amount of data needed to accurately depict the space grows exponentially with the number of features or dimensions in a dataset generally referred as the "curse of dimensionality." By lowering the number of characteristics, dimension reduction helps to reduce this problem and improves data manageability. Two common techniques for dimension reduction are Principal Component Analysis (PCA) and t-distributed Stochastic Neighbor Embedding (t-SNE).

In this section for our dataset, we employ Principal Component Analysis (PCA) for the purpose of reducing dimensions.

```
Importance of components:
                          Comp.1     Comp.2     Comp.3     Comp.4     Comp.5     Comp.6     Comp.7     Comp.8     Comp.9
Standard deviation     3.4471847 1.09569194 0.83445917 0.77126956 0.74853124 0.72570590 0.68939904 0.67535155 0.63798889
Proportion of Variance 0.5941541 0.06002704 0.03481611 0.02974284 0.02801495 0.02633245 0.02376355 0.02280499 0.02035149
Cumulative Proportion  0.5941541 0.65418116 0.68899726 0.71874010 0.74675505 0.77308750 0.79685105 0.81965604 0.84000753
                          Comp.10    Comp.11    Comp.12    Comp.13    Comp.14   Comp.15    Comp.16    Comp.17    Comp.18
Standard deviation     0.62037036 0.60234590 0.58747467 0.57019674 0.56098726 0.5579534 0.52999929 0.52371483 0.51699394
Proportion of Variance 0.01924297 0.01814103 0.01725632 0.01625622 0.01573534 0.0155656 0.01404496 0.01371386 0.01336414
Cumulative Proportion  0.85925050 0.87739153 0.89464785 0.91090407 0.92663940 0.9422050 0.95624997 0.96996383 0.98332797
                          Comp.19     Comp.20
Standard deviation     0.47973466 0.321395885
Proportion of Variance 0.01150727 0.005164766
Cumulative Proportion  0.99483523 1.000000000
```

Fig. 11: Principal Component Analysis

After applying Principal Component Analysis (PCA), we observed the following outcomes regarding the following:

Cumulative Proportion of Variance Explained by Components:

Components Overview: The analysis identified the first five principal components (Comp.1 to Comp.5) derived from PCA. These components represent the primary dimensions after reducing the dataset.

Cumulative Proportion: This metric signifies the amount of the dataset's total variance captured incrementally by each component. It reflects how much of the original data's variability is retained in the reduced dimensions.

Individual Component Contributions:

Comp.1: Alone, it captures 59.42% of the total variance in the dataset, indicating its strong influence.

Comp.2: When added to Comp.1, they cumulatively account for 65.42% of the variance. This shows an incremental but significant addition to the variance explained.

First Five Components: Cumulatively, Comp.1 through Comp.5 account for 74.68% of the total variance. This cumulative figure underscores the effectiveness of these components in encapsulating most of the dataset's variability.

Loadings:

	Comp.1	Comp.2	Comp.3	Comp.4	Comp.5
anxiety_level	0.244				
self_esteem	-0.242				
mental_health_history				-0.324	
depression	0.243				
headache				-0.266	
blood_pressure		-0.744			
sleep_quality	-0.240				
breathing_problem		0.304		0.630	-0.333
noise_level			0.277		-0.299
living_conditions				-0.361	-0.764
safety			0.343		
basic_needs			0.244	0.257	
academic_performance			0.259		
study_load			0.410	-0.284	
teacher_student_relationship			0.340		
future_career_concerns	0.248				
social_support		0.480	0.312		
peer_pressure			0.331		
extracurricular_activities			0.299	0.257	
bullying	0.243				

Fig.12: Loadings of Principal Component Analysis

Describing loading information from PCA:

Component 1 (Mental and Social Stress Factors):
Proportion of Variance: 59%
Key Variables: anxiety_level, self_esteem, depression, sleep_quality, future_career_concerns, bullying
Description: This component is heavily influenced by variables related to mental health and social stress. It seems to capture the overall mental and emotional state of individuals, including their concerns about the future and interactions with peers. High scores on this component might indicate high levels of stress and anxiety.

Component 2 (Physical Health Indicators):
Proportion of Variance: 6%
Key Variables: blood_pressure, breathing_problem, future_career_concerns
Description: Dominated by variables related to physical health, this component may represent the physiological manifestations of stress or other health conditions. The inclusion of future career concerns suggests a link between stress and physical health.

Component 3 (Environmental and Academic Influences):
Proportion of Variance: 3.4%
Key Variables: noise_level, safety, basic_needs, academic_performance, study_load, teacher_student_relationship, peer_pressure
Description: This component seems to capture the environmental factors and academic pressures that affect an individual. It includes aspects of the study environment, relationships

in the academic setting, and pressures from peers.

Component 4 (Physical and Environmental Well-being):
Proportion of Variance: 2.9%
Key Variables: mental_health_history, headache, breathing_problem, living_conditions
Description: This component might reflect the overall physical well-being and living conditions of individuals. It combines elements of mental health history with current physical symptoms and environmental factors.

Component 5 (Living Environment Quality):
Proportion of Variance: 2.8%
Key Variables: living_conditions, noise_level, breathing_problem
Description: Focused on the quality of the living environment, this component relates to how factors like noise and air quality in one's living space might impact their health, particularly respiratory health.

	Comp.1	Comp.2	Comp.3	Comp.4	anxiety_level	self_esteem	mental_health_history	depression	headache	blood_pressure
S1	-0.50	1.36	-0.19	0.62	0.48	0.25	-0.98	-0.20	-0.36	-1.42
S2	4.56	-0.18	0.14	0.26	0.65	-1.10	1.02	0.32	1.77	0.98
S3	-0.09	1.27	-0.50	-1.15	0.16	0.02	1.02	0.19	-0.36	-1.42
S4	3.69	-0.47	0.44	-0.46	0.81	-0.65	1.02	0.32	1.06	0.98
S5	-0.58	-1.46	0.05	-0.58	0.81	1.14	-0.98	-0.72	-0.36	0.98
S6	4.13	-0.15	0.85	-0.17	1.47	-0.54	1.02	1.09	0.35	0.98

	sleep_quality	breathing_problem	noise_level	living_conditions	safety	basic_needs	academic_performance	study_load
S1	-0.43	0.90	-0.49	0.43	0.19	-0.54	0.16	-0.47
S2	-1.08	0.90	0.27	-1.36	-0.53	-0.54	-1.26	1.05
S3	-0.43	-0.53	-0.49	-0.46	0.19	-0.54	-0.55	0.29
S4	-1.08	0.18	1.02	-0.46	-0.53	-0.54	-0.55	1.05
S5	1.51	-1.25	0.27	-0.46	0.90	0.15	0.87	0.29
S6	-1.08	0.90	0.27	-0.46	-0.53	-1.24	-0.55	1.81

	teacher_student_relationship	future_career_concerns	social_support	peer_pressure	extracurricular_activities	bullying
S1	0.25	0.23	0.11	0.19	0.17	-0.40
S2	-1.19	1.54	-0.84	0.89	1.58	1.56
S3	0.25	-0.42	0.11	0.19	-0.54	-0.40
S4	-1.19	0.89	-0.84	0.89	0.87	1.56
S5	-1.19	-0.42	-0.84	1.60	-1.95	1.56
S6	-0.47	1.54	-0.84	0.89	0.87	1.56

Fig.13: PCA for a few Students

From this output, selecting observation one and explaining:

Physical Health Concerns: Student1 appears to have more issues related to physical health (notably breathing problems), as indicated by the high score in Comp.2 and the individual health indicators.
Mental Health and Social Stress: While there are indications of moderate anxiety, overall mental health factors like depression and mental health history are below average, aligning with the negative score in Comp.1.
Environmental and Academic Factors: The scores and values suggest a generally positive but not outstanding environment and academic situation. The student seems to have good relationships with teachers, moderate academic performance, and some level of engagement

in extracurricular activities.

Balanced Profile: Overall, student1's profile suggests a balance between various aspects of health, social, and academic life, with a notable emphasis on physical health concerns.

CLUSTER ANALYSIS

(This section is authored by Karan Bhosale and Arshdeep Kaur)

Cluster analysis groups together similar observations to help find patterns or structures in datasets**.** It is a flexible method that may be applied to a wide range of applications and domains to find patterns, segment data, aid in decision-making, and provide new insights. Its use is especially beneficial when working with intricate datasets that have underlying structures that aren't always obvious.

For our analysis we did 2 types of cluster analysis:

1.K-Means Clustering

2.Model-Based Clustering

K-MEANS CLUSTERING

(This section is authored by Arshdeep Kaur)

In the context of K-means clustering, particularly when examining scree plots, "WGSS" stands for "Within-Group Sum of Squares". The primary goal in K-means clustering is to organize the data into a specific number of clusters, k, while minimizing the variation within each cluster, quantified by the WGSS. A lower WGSS value suggests that the data points are more tightly clustered around their respective centroids, indicating a more effective clustering.

Fig.14: Scree Plot

When analyzing a scree plot, which charts the WGSS against different numbers of clusters, we can discern the most suitable number of clusters for K-means clustering. In this scenario, the above scree plot suggests that either 4 or 5 clusters might be the optimal choice, as indicated by the WGSS trend in the plot and we chose 5 for the analysis.

```
km$tot.withinss
```

```
[1] 7197.533
```

Fig.15: Within-Group Sum of Squares of K-Means Clustering

From the analysis of within-group sum of squares (WGSS) and the scree plot visualization in our K-means clustering, it becomes evident that the choice of cluster number significantly impacts the clustering effectiveness. When opting for 4 clusters, the WGSS stands at 7560.355. In contrast, choosing 5 clusters reduces the WGSS to 7197.533, indicating a tighter grouping of data points around their respective centroids. This lower WGSS with 5 clusters suggests a more optimal clustering arrangement. Therefore, based on the lower WGSS value and the insights provided by the scree plot, selecting 5 clusters appears to be the more effective choice for our K-means clustering model.

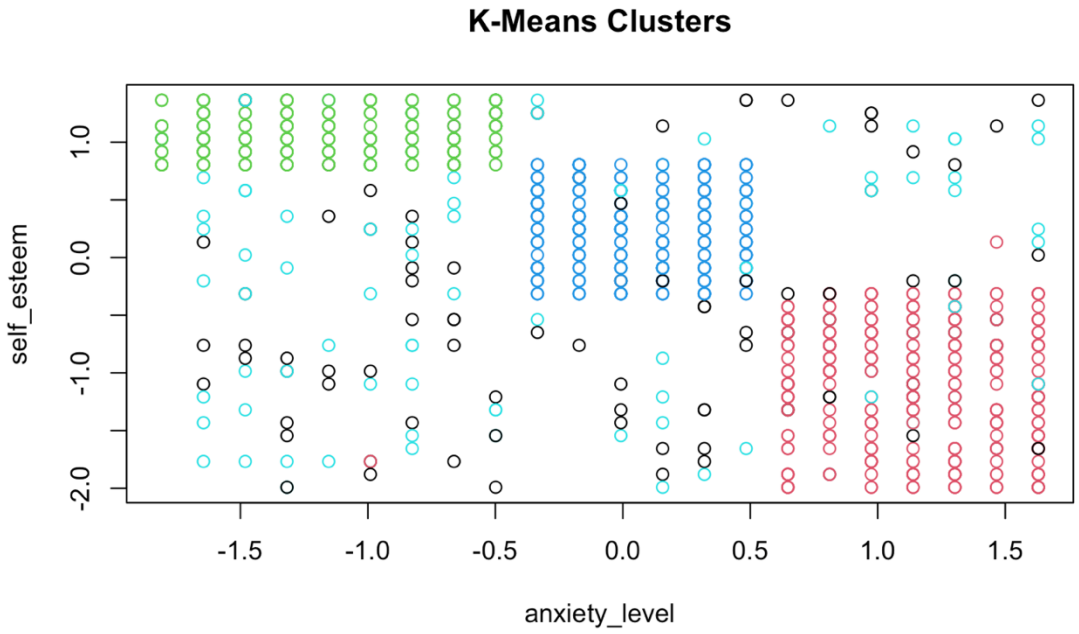


Fig.16: K-Means Clusters

This scatter plot (Fig:18) presents a somewhat mixed outcome, making it challenging to discern clear clustering patterns. However, it can be observed that the green, blue, and red clusters are distinct from each other. In contrast, the remaining two clusters appear to overlap. This observation is based on a visualization that plots the variables 'self_esteem' and 'anxiety_level' using data extracted from the original scaled dataset.

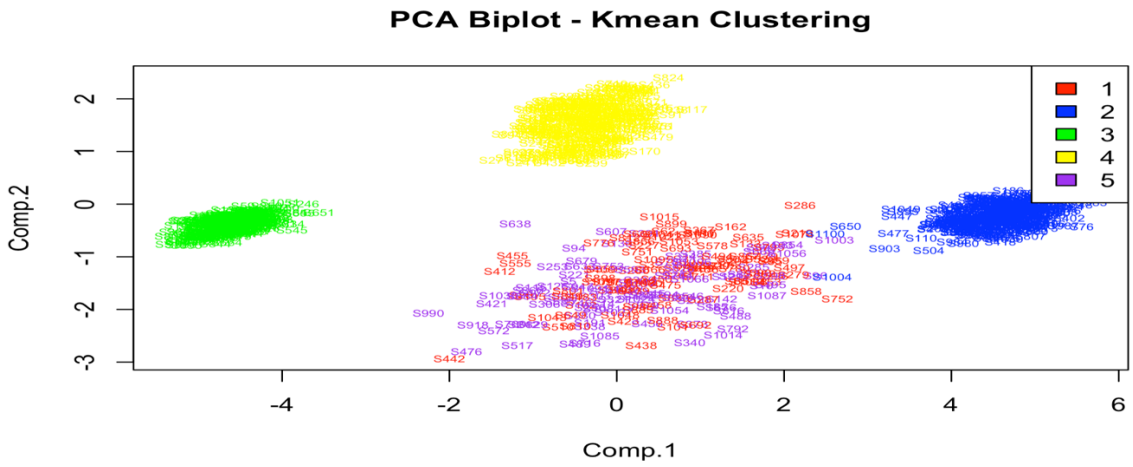


Fig.17: PCA Biplot – Kmeans Clustering

In this analysis, we implemented K-means clustering using PCA scores, choosing to form 5 clusters based on five principal components. The resulting clusters display varied degrees of separation: the blue, red, and green clusters each form distinct groupings. However, the remaining two clusters, colored yellow and purple, exhibit deep overlap.

	anxiety_level	self_esteem	mental_health_history	depression	headache	blood_pressure	sleep_quality	breathing_problem
1	-0.007484048	-0.4244891	-0.17358729	-0.1558310	-0.08008422	0.9834112	0.08604511	0.0928206
2	1.136079408	-1.1145746	1.01140912	1.1442003	1.00992222	0.9834112	-1.06946947	0.8427346
3	-1.120333389	1.0766856	-0.98143370	-1.0917550	-1.06864704	-0.2154909	1.17620006	-0.8835937
4	0.059878388	0.2445654	0.04168081	-0.1168552	-0.01433442	-1.4158534	-0.13283471	0.2028724
5	-0.296249183	-0.1526768	-0.09996903	0.3114496	0.26584344	0.9834112	0.04982811	-0.6649985

	noise_level	living_conditions	safety	basic_needs	academic_performance	study_load	teacher_student_relationship
1	-0.34878349	-0.92925745	-0.64000051	-0.02170134	-0.07668819	-0.5775202	-0.6591380
2	1.02181341	-0.88783448	-0.84762589	-0.86713621	-0.91081970	1.0242187	-0.8472290
3	-0.88573693	0.90010587	1.24919867	1.21590643	1.22301835	-0.8562719	1.3109839
4	-0.09300422	-0.03250107	-0.18790948	-0.18382513	-0.21461945	-0.1017022	-0.1215680
5	0.17786980	1.10198210	0.03992318	-0.46058568	-0.18311773	0.3467342	-0.3529365

	future_career_concerns	social_support	peer_pressure	extracurricular_activities	bullying
1	0.07463498	-1.3048834	-0.5198577	0.09585518	-0.09600735
2	1.19966397	-0.8517671	1.2330630	1.17304283	1.19146777
3	-1.07537437	1.0655852	-0.8466009	-0.89800133	-1.05342525
4	-0.12125521	0.6125296	-0.1836154	-0.19653555	-0.03463809
5	-0.15370392	-1.2218890	-0.1650601	-0.41750705	-0.29430681

Fig.18: Kmeans Centers

The output from Km$centers seems to be a summary of cluster centers obtained from a k-means clustering analysis of a dataset involving various mental, physical health, educational, social, and environmental parameters.

Here's an interpretation of the output:

Cluster Centers: Each row represents the centroid of a cluster (1 to 5) in the multidimensional space of the dataset. The values are the means of the features for all the data points within each cluster.

Features: The columns represent various features like anxiety level, self-esteem, mental health history, depression, headache, blood pressure, sleep quality, breathing problem, etc. Each feature's value at a cluster center is the average of that feature for all data points in the cluster.

Interpreting the Values:
Positive Values: Higher than average scores in the respective feature for that cluster.
Negative Values: Lower than average scores in the respective feature for that cluster.
Close to Zero: Indicates that the feature's average in that cluster is close to the overall mean of the dataset for that feature.

Cluster Characteristics:
Cluster 1: Characterized by slightly below average anxiety levels, low self-esteem, and average blood pressure. Notably, there is low social support and negative peer pressure.
Cluster 2: High anxiety level, very low self-esteem, significant mental health history, and high depression. This cluster also experiences high bullying, peer pressure, and academic performance concerns.
Cluster 3: Marked by very low anxiety, high self-esteem, and a lack of mental health history. This group has positive living conditions and safety but reports low social support.
Cluster 4: Average values in most features, indicating a balanced profile across the measured parameters.
Cluster 5: Slightly below average in self-esteem and mental health history, with average blood pressure and noise levels. This cluster has low future career concerns and social support.

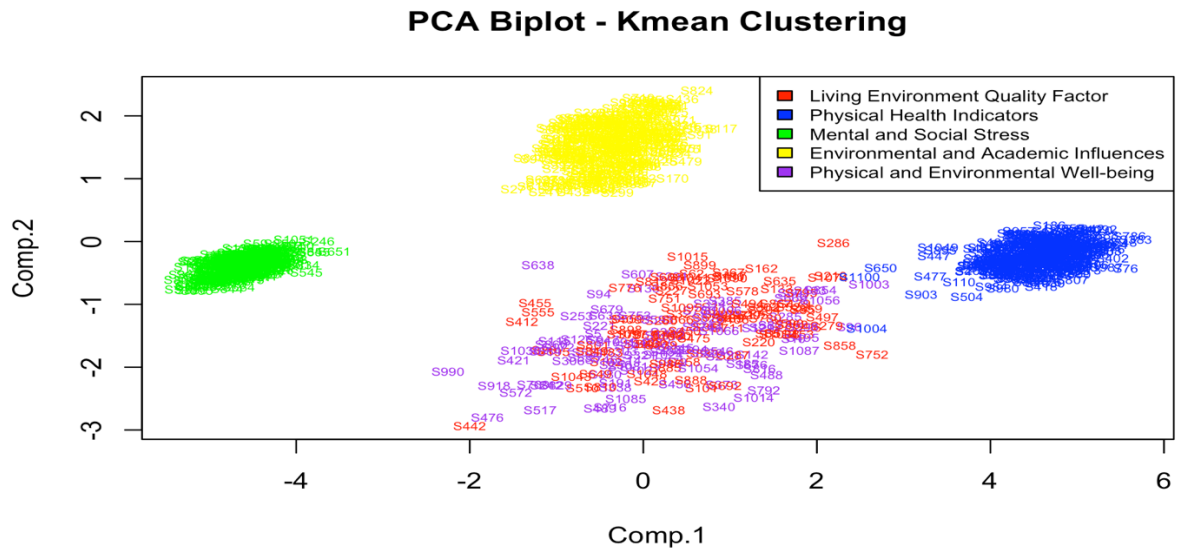


Fig.19: PCA Biplot – Kmean Clustering

Based on the K-means cluster centers and the PCA loadings, we can interpret the clusters as follows:
Cluster One is indicative of the "Living Environment Quality Factor," suggesting a focus on

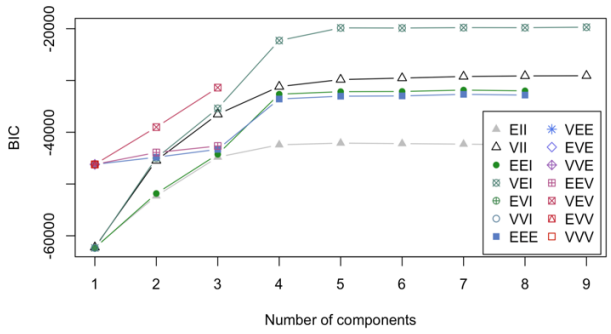
variables related to the quality of living conditions.
Cluster Two aligns with "Physical Health Indicators," pointing to elements such as physical wellbeing and health-related issues.
Cluster Three is characterized by "Mental and Social Stress Factors," highlighting aspects like mental health, stress, and social pressures.
Cluster Four is associated with "Environmental and Academic Influences," encompassing factors related to academic environments and surrounding influences.
Cluster Five corresponds to "Physical and Environmental Well-being," representing a combination of physical health and environmental quality.

MODEL BASED CLUSTERING

(This section is authored by Karan Bhosale)

Model-based clustering is an approach to cluster analysis that involves fitting statistical models to the data to identify clusters. model-based clustering assumes that the data are generated from a mixture of underlying probability distributions. The most common model-based clustering method is the Gaussian Mixture Model (GMM).

After applying model-based clustering to our dataset, we encountered a discrepancy between the theoretical expectations and the empirical results. The Bayesian Information Criterion (BIC) suggested that the optimal number of clusters is nine (as indicated by the lowest BIC value for nine clusters). However, our theoretical understanding of the dataset leads us to expect only five distinct clusters. Given this contradiction, we might consider reassessing the suitability of model-based clustering for our dataset or revisiting our theoretical assumptions about the number of clusters present.




```
Best BIC values:
             VEI,9        VEI,7        VEI,8
BIC      -19706.74 -19789.12771 -19801.23071
BIC diff      0.00    -82.38774    -94.49074
```

Fig.20: BIC plot for Model Based Cluster

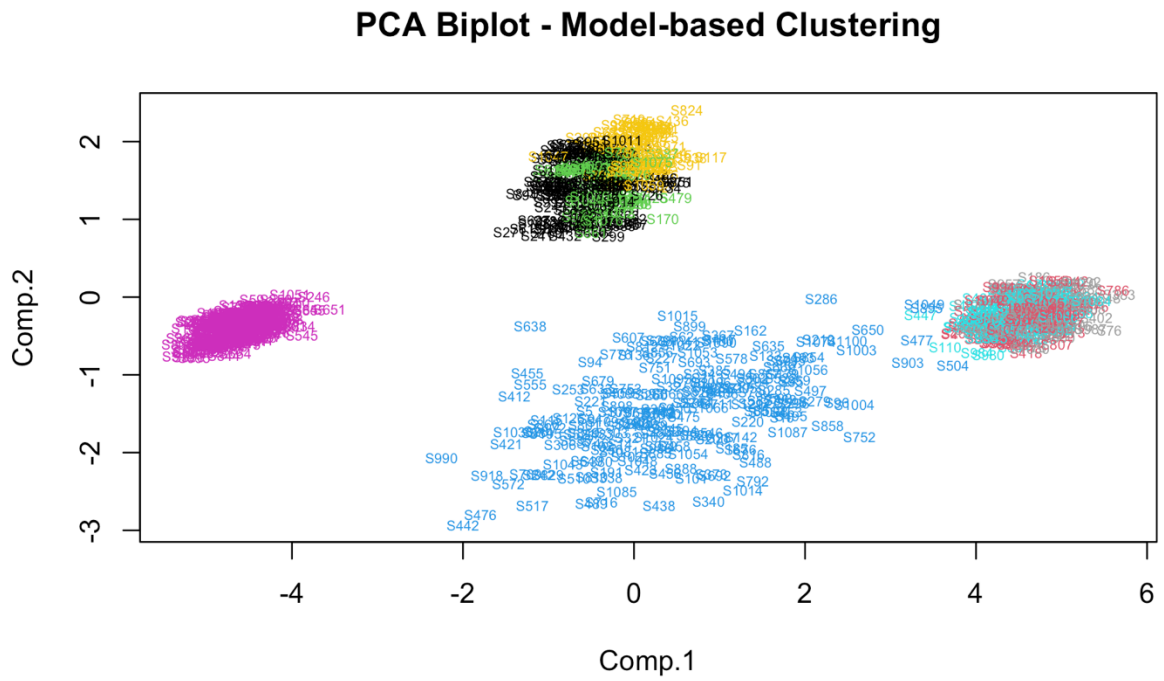


Fig.21: PCA Biplot – Model Based Clustering

Comparing K-Means and Model Based

We have determined that K-means clustering yields more effective clustering results for our dataset, leading us to choose it over model-based clustering. This decision is based on the superior performance of K-means in creating meaningful clusters, aligning with our data analysis objectives. Consequently, we have decided to proceed with the clusters identified by the K-means method.

EXPLORATORY FACTOR ANALYSIS(EFA)

(This section is authored by Karan Bhosale)

EFA is used to find latent constructs, or factors, which are inferred from patterns of correlations among observed data but may not be clearly visible. It might be difficult to understand the complexity of interactions when there are numerous interrelated variables. By locating shared elements that account for the observed correlations, EFA aids in the simplification and summary of these interactions.

Implementing Exploratory Factor Analysis (EFA) in our analysis, we observed that the approximate correlation matrix closely aligns with the actual correlation matrix, evidenced by a small Root Mean Square Error (RMSE). This indicates an ideal fit of the data for our model. Consequently, we can confidently accept our EFA model, especially since the p-value is statistically significant.

Test of the hypothesis that 4 factors are sufficient.
The chi square statistic is 213.34 on 116 degrees of freedom.
The p-value is 9.71e-08

```
rmse = sqrt(mean((corHat_stress - corr)^2))
rmse
```

```
[1] 0.0118496
```

Fig.22:

Describing loading information from EFA:

```
Loadings:
                             Factor1 Factor2 Factor3 Factor4
anxiety_level                 0.749
self_esteem                  -0.648
mental_health_history         0.662
depression                    0.713
headache                      0.650
blood_pressure                        0.975
sleep_quality                -0.742
breathing_problem             0.613
noise_level                   0.606
living_conditions            -0.581
safety                       -0.614
basic_needs                                   0.823
academic_performance         -0.647
study_load                    0.618
teacher_student_relationship -0.602
future_career_concerns        0.745
social_support                       -0.652   0.615
peer_pressure                 0.722
extracurricular_activities    0.721
bullying                      0.738
```

Fig.23: Loading of Exploratory Factor Analysis

Factor1 - "Psychosocial and Environmental Well-being":

This factor is predominantly associated with mental health and related aspects. The observed variables include:
anxiety_level (0.749)
self_esteem (-0.648)
mental_health_history (0.662)
depression (0.713)
headache (0.650)
sleep_quality (-0.742)
breathing_problem (0.613)
noise_level (0.606)
living_conditions (-0.581)
safety (-0.614)
academic_performance (-0.647)

study_load (0.618)
teacher_student_relationship (-0.602)
future_career_concerns (0.745)
peer_pressure (0.722)
extracurricular_activities (0.721)
bullying (0.738)

Factor2 - "Physical Health":
This factor is related to physical health, represented by:
blood_pressure (0.975)

Factor3 - "Social Support and Relations":
This factor is associated with social aspects, indicated by:
social_support (0.615)

Factor4 - "Basic Personal Needs":
This factor is related to basic personal needs, as represented by:
basic_needs (0.823)

In summary, these factors collectively provide a comprehensive view of different dimensions of health and well-being, each focusing on distinct but interrelated aspects: psychosocial and environmental influences, physical health, social support, and basic personal needs. The factor loadings offer insights into how strongly each variable is related to these underlying factors.

CONFIRMATORY FACTOR ANALYSIS (CFA)

(This section is authored by Priyanka Chahande)

Confirmatory Factor Analysis is a method for testing and validating the underlying structure of a set of observed variables and assessing how well the observed variables represent latent constructs or factors.

CFA is applied to the student stress dataset to analyse how well the manifest variables represent the latent factors. Our theory was that there are five latent factors. The variables anxiety_level, self_esteem, mental_health_history and depression are driven by psychological_Factors. The variables headache, blood_pressure, sleep_quality, and breathing_problem are driven by physiological factors. The variables noise_level, living_conditions, safety, and basic_needs are driven by environmental factors. The variable academic_performance, study_load, teacher_student_relationship, and future_career_concerns are driven by academic factors. Lastly, the variables social_support, peer_pressure, extracurricular_activities, and bullying are driven by social factors.

While modeling the data using the SEM package, the model did not converge and errored out.

```r
opt <- options(fit.indices = c("GFI", "AGFI", "SRMR")) # Some fit indices
sim_sem2 <- sem(sim_model, cov(data_dim[-1]), nrow(data_dim[-1]))
summary(sim_sem2)
```

```
Error in summary.objectiveML(sim_sem2) :
 coefficient covariances cannot be computed
```

Show Traceback

We can conclude that our data is not ideally suited for the Confirmatory Factor Analysis (CFA) model. Therefore, we have decided to exclude our data from this modeling approach.

Fig.24: SEM Model Error

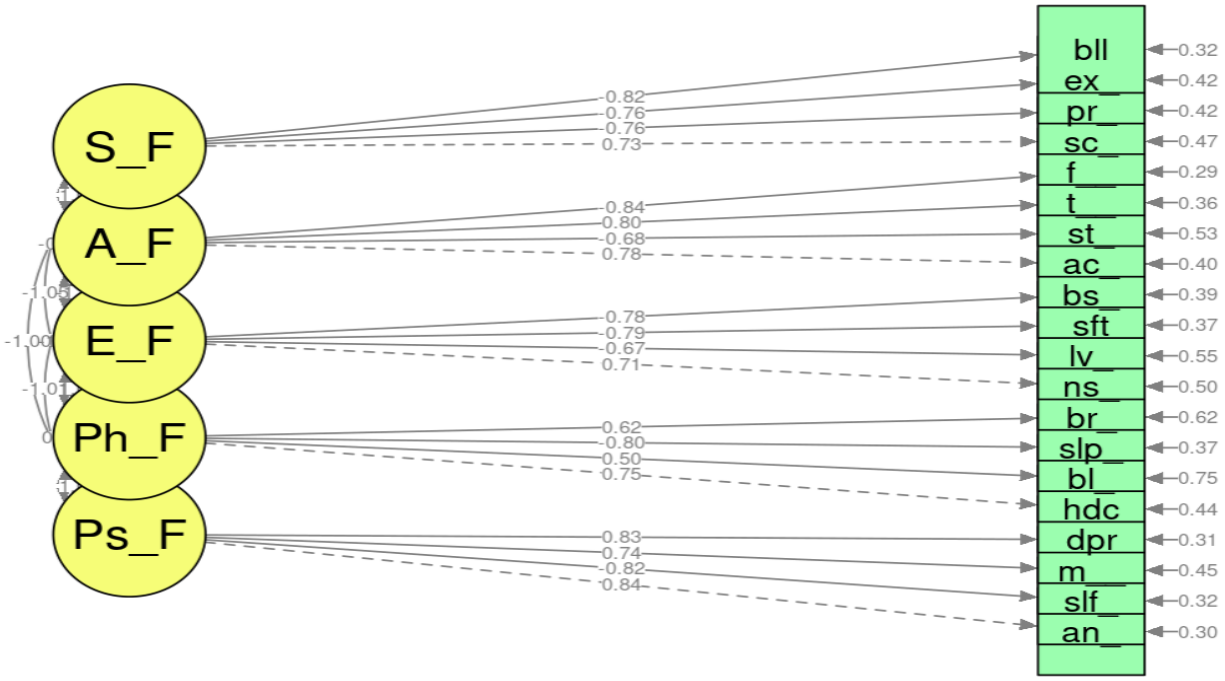


Fig.25: Path Diagram for CFA Model

Lavaan package was then used to build the CFA model. The above figure (Fig: 17) shows path diagram for the CFA model.

```
Model Test User Model:

  Test statistic                              1775.272
  Degrees of freedom                           160
  P-value (Chi-square)                         0.000
```

Fig.26: Output of CFA Model

However, the CFA model shows that P-value < 0.05, which implies that the CFA model does not support the data.

REGRESSION ANALYSIS

(This section is authored by Nahid Ferdous)

Linear regression (Used lmtest library)

In the regression analysis we're conducting, the PCA scores are being utilized as predictor variables, while the stress level variable from our actual dataset serves as the response variable. This method leverages the principal components derived from PCA to predict the stress levels. The principal components are essentially a transformation of our original variables into a new set of uncorrelated features, summarizing key patterns in the data. By using 5 principal components as predictors, the regression model aims to understand how these underlying patterns relate to the stress levels in our dataset.

```
Call:
lm(formula = stress_level ~ ., data = pca_scores_df)

Residuals:
     Min       1Q   Median       3Q      Max
-1.63469 -0.06390  0.01194  0.08979  1.37216

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.993601   0.011638  85.378  < 2e-16 ***
Comp.1       0.209749   0.003376  62.129  < 2e-16 ***
Comp.2       0.034890   0.010621   3.285  0.00105 **
Comp.3       0.022974   0.013946   1.647  0.09979 .
Comp.4      -0.034336   0.015089  -2.276  0.02307 *
Comp.5      -0.041330   0.015547  -2.658  0.00797 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3849 on 1088 degrees of freedom
Multiple R-squared:  0.7813,    Adjusted R-squared:  0.7802
F-statistic: 777.2 on 5 and 1088 DF,  p-value: < 2.2e-16
```

Fig.27: Output of Linear Regression

Short summery of the regression analysis:

Residuals: The residuals (the differences between observed and predicted values) seem reasonably distributed around 0, which is a good sign.

Multiple R-squared: 0.7813 indicates that approximately 78.13% of the variability in stress levels is explained by our model. This is a strong model in terms of explanatory power.

Adjusted R-squared: 0.7802 adjusts for the number of predictors in the model and is also relatively high, indicating a good fit.

F-statistic: The F-statistic tests the overall significance of the model. A very low p-value (< 2.2e-16) suggests that our model is statistically significant.

CONCLUSION

(This section is authored by Priyanka Chahande and Arshdeep Kaur)

Following are the key findings and conclusions from the above analysis:
Data Cleaning and Visualization: To guarantee data accuracy and to spot trends, the first phase requires data cleaning and visualization. Mean imputation was used to fill in the missing values, and the dataset was presented to help with understanding of the distributions, types of variables, and outlier presence.

Dimension reduction analysis: To reduce dimensions, Principal Component Analysis (PCA) was used. The first five main components were found through analysis, and together they accounted for 74.68% of the variation. The underlying variables impacting the well-being of students were considered while interpreting these components.

Model-Based and K-Means Cluster Analysis: Five groups were chosen for K-Means clustering based on the examination of scree plots and the within-group sum of squares (WGSS). The various features pertaining to mental, physical, environmental, and academic variables were used to interpret the clusters.

Based on the Bayesian Information Criterion (BIC), model-based clustering proposed nine clusters; however, the practical interpretation prompted a reevaluation, and K-Means clustering was found to be more efficient.

Exploratory Factor Analysis: To find latent components or factors impacting students' well-being, exploratory factor analysis, or EFA, was used. Four factors were identified by the analysis: physical health, social support and relationships, psychosocial and environmental well-being, and basic human needs.

Regression Analysis: Stress levels served as the response variable, and PCA scores were used as predictors in a regression analysis. The model demonstrated a high Multiple R-squared (0.7813), meaning it could account for 78.13% of the variation in stress levels.

In conclusion, this project has not only identified the key factors influencing student stress levels but has also provided a robust analytical framework for understanding and interpreting these influences. The methodologies and findings of this analysis could serve as valuable inputs for educational institutions, policymakers, and student support services in designing targeted interventions and support mechanisms to enhance student well-being and reduce stress. This project stands as a testament to the power of multivariate data analysis in unraveling complex real-world issues and providing actionable insights.

STRATEGIES RECOMMENDED BASED ON ANALYSIS

(This section is authored by Arshdeep Kaur and Karan Bhosale)

Following the analysis of stresses and student well-being, the following coping mechanisms can be suggested to address the concerns that were found:

Individualized Support Programs: Support programs according to the factors and clusters that have been identified should be developed. Acknowledge that students may require different kinds of support, and that individualized care can address issues with mental, physical, social, and academic well-being.

Mental Health Resources: Improve the counseling services, workshops, and awareness campaigns for mental health that are available on campus. Interventions should be targeted according to the psychosocial and mental health well-being determinants, such as social support, anxiety, depression, and self-esteem.

Initiatives for Physical Health: Encourage physical health by implementing wellness and health initiatives. Offer tools and exercises to treat blood pressure, respiration, and other associated disorders.

Social Support Networks: Initiatives by fostering a sense of community and belonging that promote healthy social interactions, extracurricular activities, and peer mentoring programs should be encouraged.

Encouragement of Self-Care Activities: Students should be encouraged to exercise self-care. Offer tools and data regarding self-care practices, mindfulness, and stress reduction. Give pupils the abilities to take charge of their own well-being.

Talking and Being Aware: Encourage honest discussion about challenges pertaining to well-being and the resources for assistance that are available. Decrease the stigma attached to mental health issues and spread knowledge about how important it is to get treatment.

REFERENCES

[1] https://www.kaggle.com/datasets/rxnach/student-stress-factors-a-comprehensive-analysis/data

[2] The_psychometric_properties_of_the_Generalized_Anx.pdf

[3] chat.openai.com