Problem Statement : Stress level prediction or Mental health status of undergraduate students.

- This prediction is based on the amount of sleep the person gets, the CGPA of the person, his/her interest in the field of study and the amount/ frequency of sports engagement.
- This data set will be helpful for researchers as well as the college administrators as it will predict the mental health of student and hence, the administrators can find ways to mitigate the stress level problem.

Domain : Healthcare and Welfare.

Dataset Involves (Structured data): Mental Health survey data. (Its main focus is on students studying data science, computer science and software engineering, but to avoid vastness of data these attributes were ignored)

Source of original dataset (Secondary data collection): https://www.kaggle.com/datasets/abdullahashfaqvirk/student-mental-health-survey

- Excel file link (contains original and transformed dataset) : https://docs.google.com/spreadsheets/d/1-4dwnw1ZslAblMb2B0MdYSU32Xchllil/edit?usp=sharing&ouid=109195069563795767118&rtpof=true&sd=true

Context : This data includes various mental health indicators.

- The original dataset contains 87 rows and 21 columns.
- 2 rows containing postgraduate students records were deleted as the problem formulation is only about undergraduate students.
- Attributes were renamed as per convenience.
- Missing values in CGPA attribute were replaced by median CGPA group. ((0.0-0.0) changed to (2.5-3.0))
- Attributes named as gender, age, university name, degree level, residential status, campus discrimination, academic workload, social relationships and stress relief activities were discarded due to irrelevancy and ambiguity.
- For correlation of nominal data (Chi-Square Test), “Sports engagement” column was tested against 3 levels of stress namely, low, medium and high.
- For co-variance of numeric data, “study satisfaction” column was tested against total stress (sum of all kinds of stress).

The following is the description of those columns (attributes) after transformation process :

1. CGPA (Qualitative / Categorical) (Ordinal):
This column represents the CGPA (Cumulative Grade Point Average) of a student. Its been divided in categories or groups and it is in 4 point CGPA system.

2. Sports engagement (Qualitative / Categorical):
This column contains data on the amount of engagement a person has to sports. Its categorical in nature.

3. Average sleep (Qualitative / Categorical):
This column contains data on the average sleep a student gets. It is also categorical in nature, but can be converted to numeric if needed.

4. Study satisfaction (Quantitative / Numeric (Discrete)) :
This column relates the interest or satisfaction of student about the study. Its numeric and discrete. Its minimum value is 1 which means low satisfaction for student and maximum value is 5 which means high satisfaction for student.

5. Total mental stress (Quantitative / Numeric (Discrete)) (Range: 5 to 30) :
In the original dataset, stress has been showcased in form of various parameters, namely academic stress, financial stress, etc. But, for the sake of our argument, “Total stress”, which is the sum of all stress is considered for computation.
Other ways of dealing with this is to have weights given to each kind of stress. One might like to add more preference of academic stress than other stresses, hence he/she can get a total stress value which will be sum of product of weights and its respect stresses.

Preview of few rows of the dataset :

Sr. no.	cgpa	Sports engagement	average sleep	study satisfaction	academic pressure	financial concerns	depression	anxiety	isolation	future insecurity	Total mental stress	Total mental stress (categories)	Helper Column (study satisfaction * Total mental stress)
1	3.0-3.5	No Sports	4-6 hrs	5	5	4	2	1	1	2	15	medium	75
2	2.5-3.0	1-3 times	2-4 hrs	5	5	3	2	3	3	1	17	medium	85
3	2.5-3.0	No Sports	4-6 hrs	3	4	4	5	5	5	3	26	high	78
4	3.0-3.5	No Sports	4-6 hrs	3	5	2	5	5	4	4	25	high	75
5	3.0-3.5	No Sports	4-6 hrs	4	5	3	5	5	5	5	28	high	112
6	3.0-3.5	No Sports	4-6 hrs	3	4	5	3	2	2	4	20	medium	60
7	2.5-3.0	1-3 times	4-6 hrs	3	3	4	3	4	3	5	22	high	66
8	2.5-3.0	No Sports	4-6 hrs	3	3	5	5	5	5	5	28	high	84

Correlation test of nominal data (Chi-Square Test)

	Low Stress		Medium Stress		High Stress		Total
No sports	5	5.43529411764706	15	19.2705882352941	22	17.2941176470588	42
1-3 times	1	2.84705882352941	14	10.0941176470588	7	9.05882352941176	22
4-6 times	4	1.42352941176471	6	5.04705882352941	1	4.52941176470588	11
7+ times	1	1.29411764705882	4	4.58823529411765	5	4.11764705882353	10
Total	11		39		35		85

values	Column 1	Column 2	Column 3
Row 1	0.0348612172141586	0.946412411118286	1.28051220488197
Row 2	1.1982984929509	1.51136706430826	0.467914438502672
Row 3	4.66319883325228	0.179925956396545	2.7501909854851
Row 4	0.0668449197860948	0.0754147812971349	0.189075630252101

Chi squared	13.3640169354455
Degree of freedom	((4-1)*(3-1)) = 6

- Null hypothesis : There is no relation between amount of sports played and stress levels of a person.
- Alternative hypothesis : There is a relation between amount of sports played and stress levels of a person.

Considering P=0.05,

As 13.36 (chi squared value) > 12.592,

Hence, Null hypothesis is rejected and Alternate hypothesis is accepted.

Hence, there is a relation between amount of sports played and stress levels of a person.

Covariance for numeric data

Sum of all values in study satisfaction	333	Number of observations	85
Mean of values in study satisfaction (A')	3.91764705882353		
Sum of all values in total mental stress	1683	Number of observations	85
Mean of values in total mental stress (B')	19.8		

Covariance = E(A.B) - A'B'	
E(A.B) =	75.6470588235294
Covariance	-1.92235294117649

Hence, as covariance is negative, both study satisfaction and mental stress are inversely correlated. That is, if one increases then other decreases and vice-versa.