Student stress factors

Pham Quang Bach

April 14, 2024

Chapter 1

Introduction

The aim of this project is to determine the relationship between Engineering student's level of stress and various life style and social factors.
The research question that this project attempt to answer is as follows "What are the common habits or environments of students with high or low level of stress?".
The project is motivated by the personal desire to help university student deal with stress by attempting to identify factors that could contribute to stress.

Multiple Correspondence Analysis (MCA) will be use to conduct statistical analysis in this project. MCA is chosen because it could describe the dependencies between the variables, and because it would work well on the dataset consist of categorical data.

Chapter 2

Description of the Dataset

The dataset used is the Student stress factors[1] dataset, containing 520 instances of 6 different metrics on various aspect of the student's life that could be contribute to tress factor:

- Sleep quality(SQ).
- Headaches(HE).
- Academic performance(AP).
- Study load(SL).
- Extracurricular activities(EA).
- Stress levels(ST).

All data are collected primarily on University and College level Engineering students with online surveys. All metrics are categorical with categories ranging from 1 to 5. So no more pre-processing of the data are needed for the analysis.

Chapter 3

Presentation of the variables

In both univariate and bivariate statistical analysis in this chapter, the categorical data (from 1 to 5) will be treated as numerical data with the same value as the categories. I believe this is reasonable as the categories all represents intensity of the metrics and are reasonably linear.

3.1 Univariate analysis

The figures below depicts the dataset as bar plots as well as the mean, median and standard deviation of the data in each metrics.

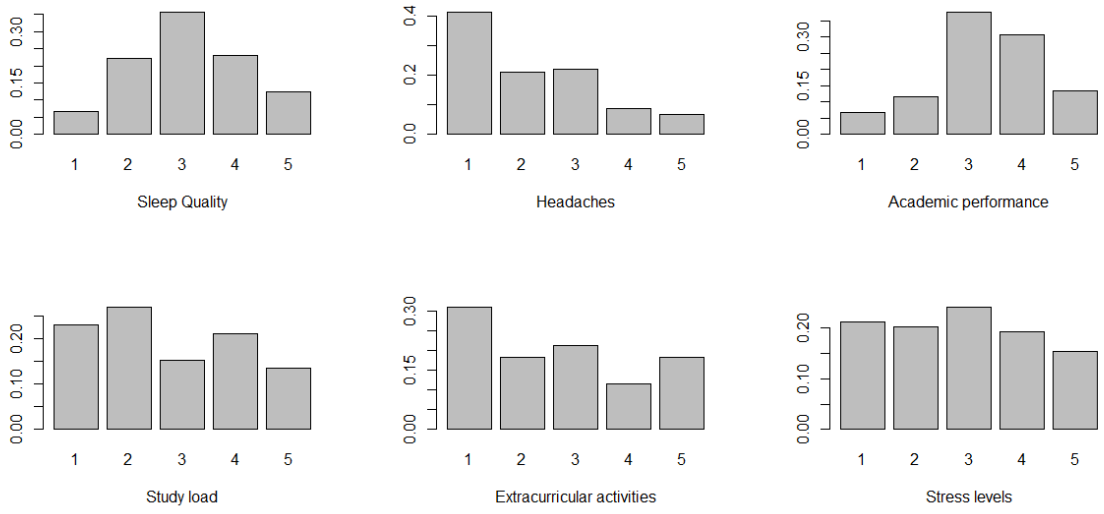

Figure 3.1: Bar plot of the data

	Mean	Median	Standard deviation
SQ	3.12	3	1.1
HE	2.18	2	1.25
AP	3.33	3	1.06
SL	2.75	2.5	1.37
EA	2.68	3	1.47
ST	2.88	3	1.36

Figure 3.2: Mean, Median and Standard deviation of the data

Looking at the bar plots, it is clear that in Study load, Extracurricular activities and Stress levels in particular, there are large deviations from the expected mean. Which is worth looking into.

3.2 Bivariate analysis

For bivariate analysis, a correlation matrix is used to examine the correlation between the metrics.

Figure 3.3: Correlation heatmap of the metrics

Looking at the heatmap, it can be seen that there is a clear correlation between study loads and the level of stress. This is expected to some extent, since most student's first priority and source of stress lies in their school work. What is not expected, however, is a slight correlation between sleep quality and stress at 0.17.

Another expected correlation are between academic performance and sleep quality, and the negative correlation between headaches and academic performance. It is quite clear why these correlations are expected, a student who has trouble sleeping and headaches is more likely to have lower academic performances.

Chapter 4

Multiple Correspondence Analysis

4.1 Analysis

In this project, using R, the method mjca from the package ca was used to perform the multiple correspondence analysis on the dataset.

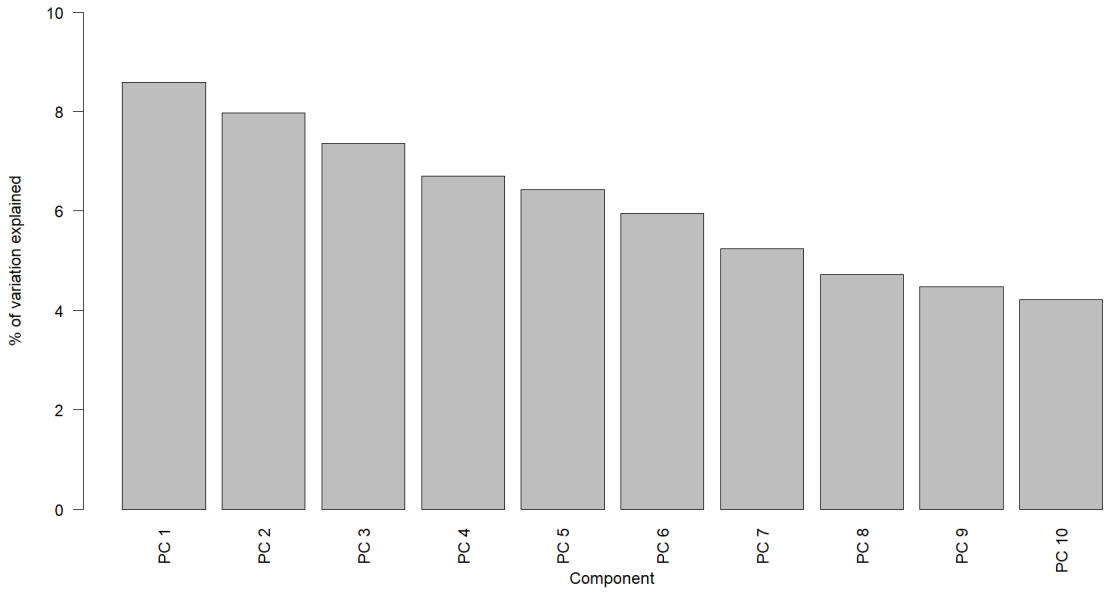

Figure 4.1: Percentage of variance explained the first 10 components

Based on the bar plot, the first 4 components only explains 30.6% of the variation

in the dataset. And there are no clear drop in contribution of the components even up to the tenth. However, because of limitations in this project and my understanding, further analysis will be focused only on the first 4 components of the MCA.

Plotting the column scores with the first 2 dimension results in the following graph.

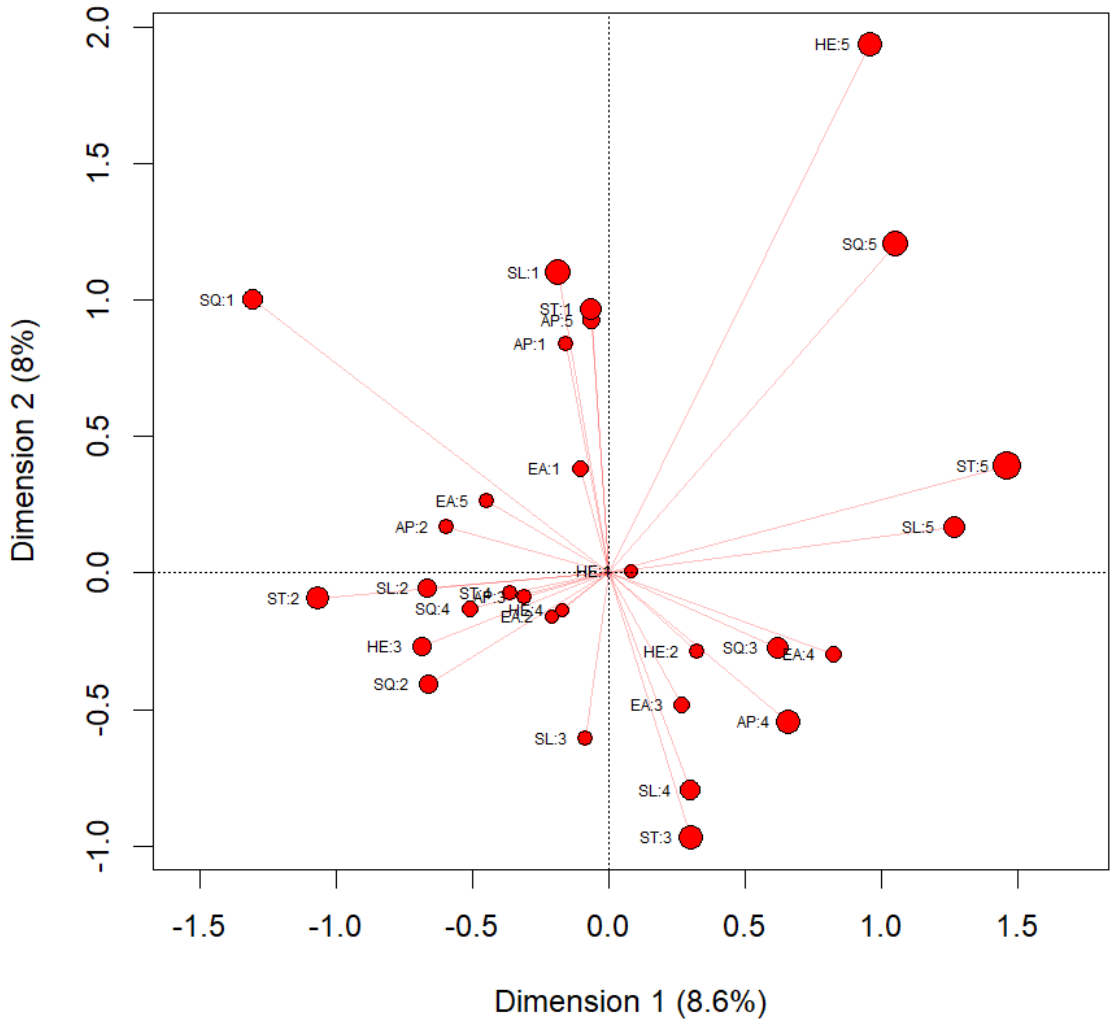

Figure 4.2: First two column principal coordinates.

Where the angle between two modalities represents attraction and repulsion. Two modalities attracts if the angle is less than 90 degrees, and more than 90 degrees means repulsion.

From the graph, it can be seen that students who have a low study load(SL:1) also have a low stress level(ST:1) and high stress level(ST:5) also attracts high study loads(SL:5). This is consistent with the results of the correlation analysis previously. Interestingly, low study load and stress level also has attraction to both low academic performance(AP:1) and high academic performance(AP:5), but high academic performance has a much higher quality of representation in this graph.

On another note, it is seen that students with below average sleep quality(SQ:2) also tends to have headaches(HE:3). Perhaps this could be explained by the sleeping habits of the students, it is logical that better sleep would results in less headaches.

Plotting the column scores with dimensions 3 and 4 results in the following graph.

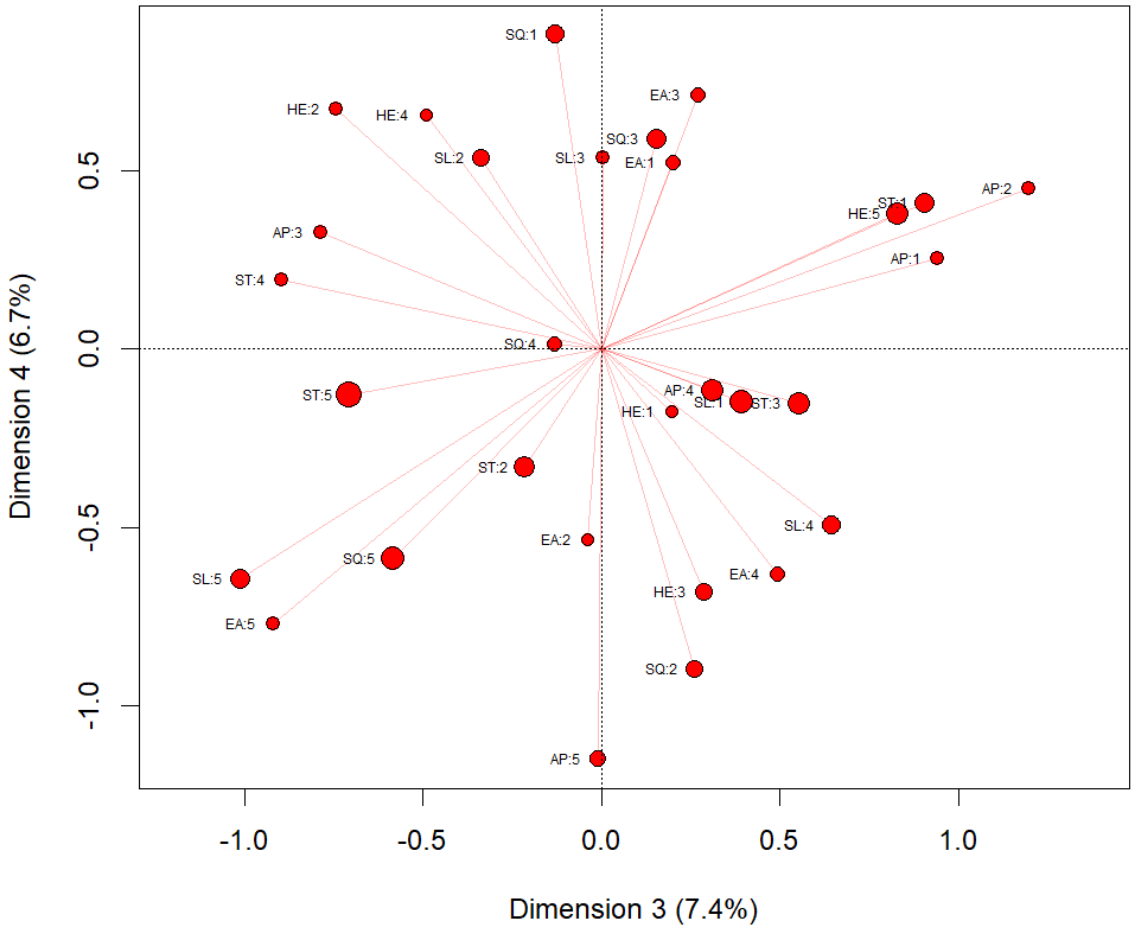

Figure 4.3: Second two column principal coordinates.

From this graph, It seems that students who have a high study load(SL:5) also have a high sleep quality(SQ:5) and less than average stress level(ST:2). This is an unexpected but not entirely surprising result, as better sleep can reduce stress. However, I personally cannot explain the correlation between study load and sleep quality or it's connection to less than average stress.

Another interesting observation is that students who have many headaches(HE:5) also have a low stress level(ST:1) and low academic performance(AP:2, AP:1). In contrast,

students how have higher academic performance(AP:4) also has lower study load(SL:1) and average stress level(ST:3). This, however, is in contrast to the findings in the previous graph, which shows that categories SL:1 and ST:3 have high repulsion, being almost directly opposite of each other. I personally do not believe I can adequately explain this discrepancy due to a lack of knowledge on both the model and method.

4.2 Conclusion

Using data found in the previous analysis, we can now answer the research question posed from the beginning. For students with low level of stress, they also have more headaches, but a lower study load. On the other hand, for students with high level of stress, they also have a high study load. From this dataset, it seems the most important and consistent habit or environmental factors of student's level of stress is the study load that the student is under.

This, however, doesn't mean that the study loads is responsible for the stress level present in the dataset. It is more likely that stress is cause by a combination of complicated factors, that was not accounted for in the dataset.

Chapter 5

Evaluation of the analysis

The analysis performed above is susceptible by multiple form of bias. One source of bias comes form the dataset itself, as the data was collected with online surveys from universities in Mumbai. Because the dataset was not collected for every student, but only students who answer the survey, this introduce a bias for the students included in the study.

Another inaccuracy could be contributed to the method used. Since MCA is a quite non-robust method, rare modalities have a big impact on analysis. In this case, several rare modalities exist within the dataset. For example, HE:5, AP:1 and SQ:1 are all quite rare. I am personally unaware on how much this impacted the analysis but it is a concern.

Finally, more bias could be attributed to my bias towards the subject and my lack of knowledge about both the topic of student stress and the method used, leading to

possible misleading or misinterpretation of the results.

References

[1] https://www.kaggle.com/datasets/samyakb/student-stress-factors