Statistical and Visual Analysis of Surveys to Identify Risk Factors for Poor Mental Health Among IT Students in Pakistan

Vic Permakoff
CS 439 Intro to Data Science, Section 03
Professor Naina Chaturvedi

Ia) Introduction

Mental health has become something that is widely regarded as an essential part of any person's ability to thrive. As young adults, many of whom are learning to live on their own, determine their identities, and begin their careers (among many other pressures), university students may especially struggle with poor mental health (Eisenberg, Hunt, & Speer, 2013). Fortunately over the past few decades, there has been less and less stigma to seek out help, but nonetheless, mental health is still something that is heavily underlooked in many parts of the world. In non-Western countries, there tends to be a lack of research on mental health and how it could affect people's lives. This lack of interventions for individuals who suffer from poor mental health (including psychological disorders such as Major Depressive Disorder and different anxiety disorders) could lead to further difficulty with everyday functioning, the ability to uphold social relationships, and one's internal self-esteem. Thus it is crucial that there exists data and proper data analysis to learn how to prevent poor mental health in such areas.

The dataset that I chose to investigate is a collection of responses to a mental health survey among Internet Technology (IT) students in Pakistan. Data on mental health among Pakistani students is relatively lacking compared to that of students from Western countries. It is known that mental health complications can exacerbate difficulties with learning and prevent academic growth; this is something that has been found to be true for Pakistani students as well (Zada et al., 2021). Furthermore, there is a significant gap between those who experience mental health problems and those who receive treatment in Pakistan (**[CITE]**). This may be in part due to the relative lack of mental health literacy among the Pakistani public, many of whom may attribute mental health to be *caused* by stress, poor emotional regulation skills, loneliness, and other factors, including many of which have superstitial or cultural significance **(**Shafiq, 2020**)**. In addition, I could not find much data about students who major in IT and similar fields, suggesting another gap in the literature. Therefore it is important that more mental health data is collected and properly analyzed from populations that are understudied such as Pakistani IT students so that interventions could be created, risk factors could be assessed, and students with poor mental health are given the help they need.

In the field of psychology, data analysis is an essential part of conducting research. Finding correlations between different parts of an individual's life could be what leads to the development of new mental health interventions. As such, I chose to use methods in line with an unsupervised learning project, mainly data visualization and statistical analysis, as I believe these techniques accurately represent real-life methods used in psychology research.

In addition, I also wanted to try using some of the statistical tests I have learned in other classes, such as t-tests, to find whether certain correlations between variables were significant. For this project, I posed a hypothesis that Pakistani IT students with high levels of mental distress (such as depression and anxiety) are more likely to be less socially involved and have lower trends of academic success. The null hypothesis would be that Pakistani IT students who have high levels of mental distress do not differ from those who have low levels of mental distress in their social involvement and academic success.

Ib) Personal Motivation

As a psychology and computer science double major, I found this project to be an excellent way for me to combine my interests and practice data analysis. After graduating I am hoping to do research in psychology, so I believe the skills I learned and the methods I used in this class (and especially for this project) will prove helpful to me in the future, especially data visualization and statistical analysis techniques such as linear regression.

II) Method and Results

Searching through the open datasets website Kaggle, I found a dataset titled "Student Mental Health Survey" containing information about the mental health and related variables of IT students from the Punjab University College of Information Technology (PUCIT) in Lahore, Punjab, Pakistan. The dataset consists of 87 participants (rows) and 21 survey questions (columns). All of the data is quantitative (close-ended, rather than open-ended) and discrete. Some of the columns contain numerical data (of which all except the "age" column were ordinal, measured on a scale of 1 through 5), while others are categorical. A more detailed description of the exact information collected can be found in the Jupyter notebook under "Column Interpretations."

The dataset was first cleaned for preprocessing by removing any NaN values (of which none were found, allowing for the data from all 87 participants to be used). The data was then analyzed in the "Column Interpretations" sections, where I also looked over the means of the numerical data to check for outliers, as well as to get an idea of trends in the data. For instance, the means and medians for variables such as "academic_workload" and "academic_pressure," which measured the subjective amount of work and pressure that the participant felt, seemed relatively high–both variables had medians of 4.0 (out of a scale of 1-5), and they had means of 3.931 and 3.885 respectively. For academic workload, no students reported having the minimal amount of workload (min=2). Furthermore, the means for the variables "depression" and "anxiety" are both around 3.2. Assuming that the scaling for these variables would follow a normal distribution, for which the expected mean would be a 3.0, this is slightly above the expected average. However, it is also difficult to draw any conclusions from this information alone, as the ranking is entirely subjective, and my hypothesis is reliant on finding a correlation between variables.

A more effective way to analyze the dataset is through visualization of relationships between variables. To visualize these relationships, I created several graphs using Seaborn and Matplotlib. First, I used a heatmap (Figure 1) to find correlations between the ordinal variables including "study_satisfaction," "academic_workload," "academic_pressure," "financial_concerns," "social_relationships," "depression," "anxiety," "isolation," and "future_insecurity." This was extremely useful in helping me see which variables I should further analyze to have a possible significant correlation. For instance, the two variables with the highest correlation are depression and anxiety (0.84); this makes sense, as the two disorders are highly comorbid and commonly found to coexist in individuals with mental health problems (Pollack, 2005). Similarly, the inverse correlation between isolation and social relationships (-0.56) also makes sense, as individuals who isolate themselves or feel isolated from others would be less likely to feel content about their level of relatedness with others.

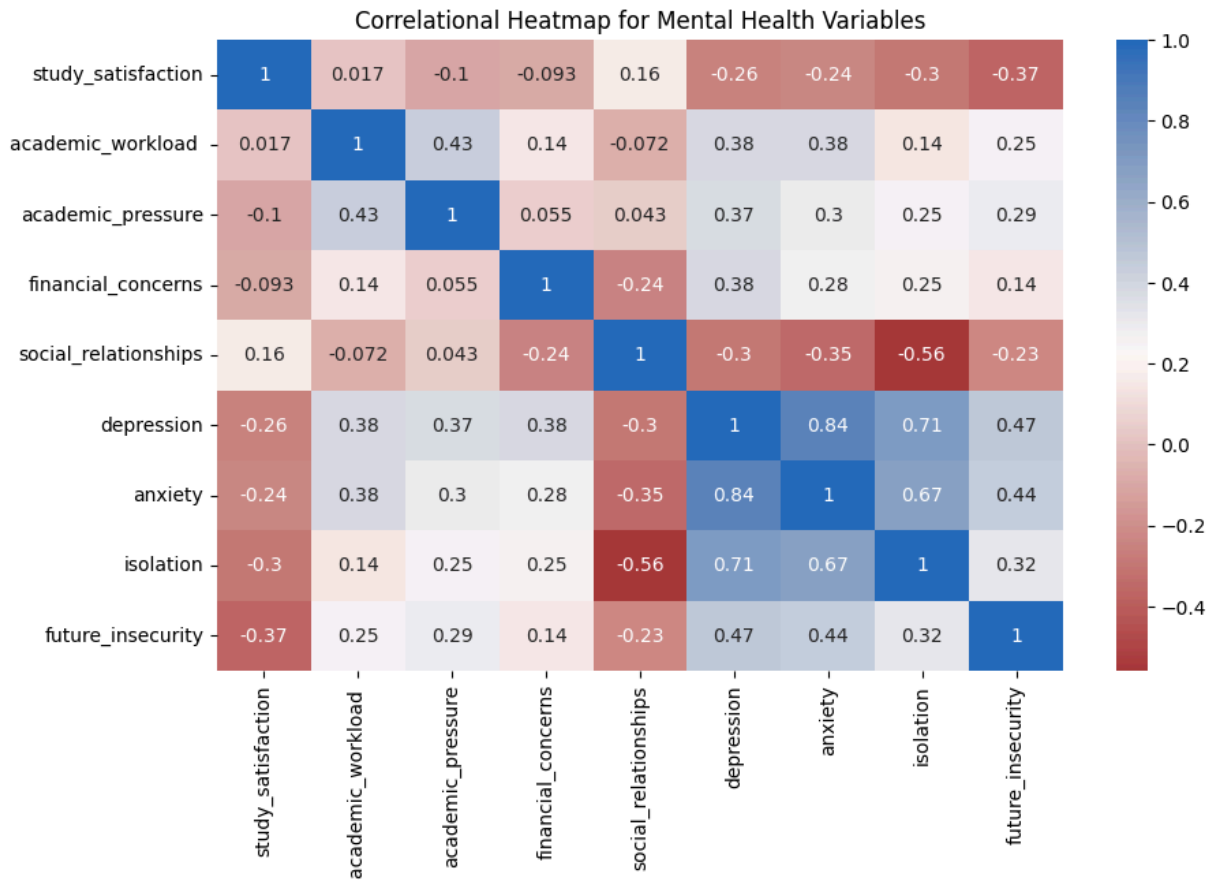

Figure 1) Heatmap for ordinal variables

Some correlations that I found more interesting and worthy of further analysis were depression/anxiety and isolation, depression/anxiety and future insecurity, depression/anxiety and social relationships, and depression/anxiety and study-satisfaction. To examine these in more detail, I first created a linear regression plot to more clearly depict the high correlation anxiety and depression. After this, I created a new categorical column in the dataset, "high_depanx" (meaning higher than average levels of depression and anxiety) in place of the depression and anxiety columns, which listed whether the individual had a high level of depression, anxiety, neither, or both. This would allow me to look at exact relationships between common mental disorders and other values, while also noticing the differences between those who have anxiety and those who have depression.

After this, I created a few histograms (Figure 2, Figure 3) with kernel density estimates to analyze the relationship between poor mental health and various variables. Of most interest to me were the variables "isolation" and "social_relationships," as these variables were most representative of what I was looking for in my hypothesis. As expected, individuals with higher rates of depression and anxiety also tended to report higher levels of isolation. This is especially significant for individuals who live off-campus, as it seems they tend to have higher levels of isolation than those who live on-campus. Interestingly, it seemed like individuals who lived off-campus also experienced higher levels of campus discrimination; however, the proportion of students who live on-campus is very small compared to those who live off-campus, making it hard to draw a definite conclusion.

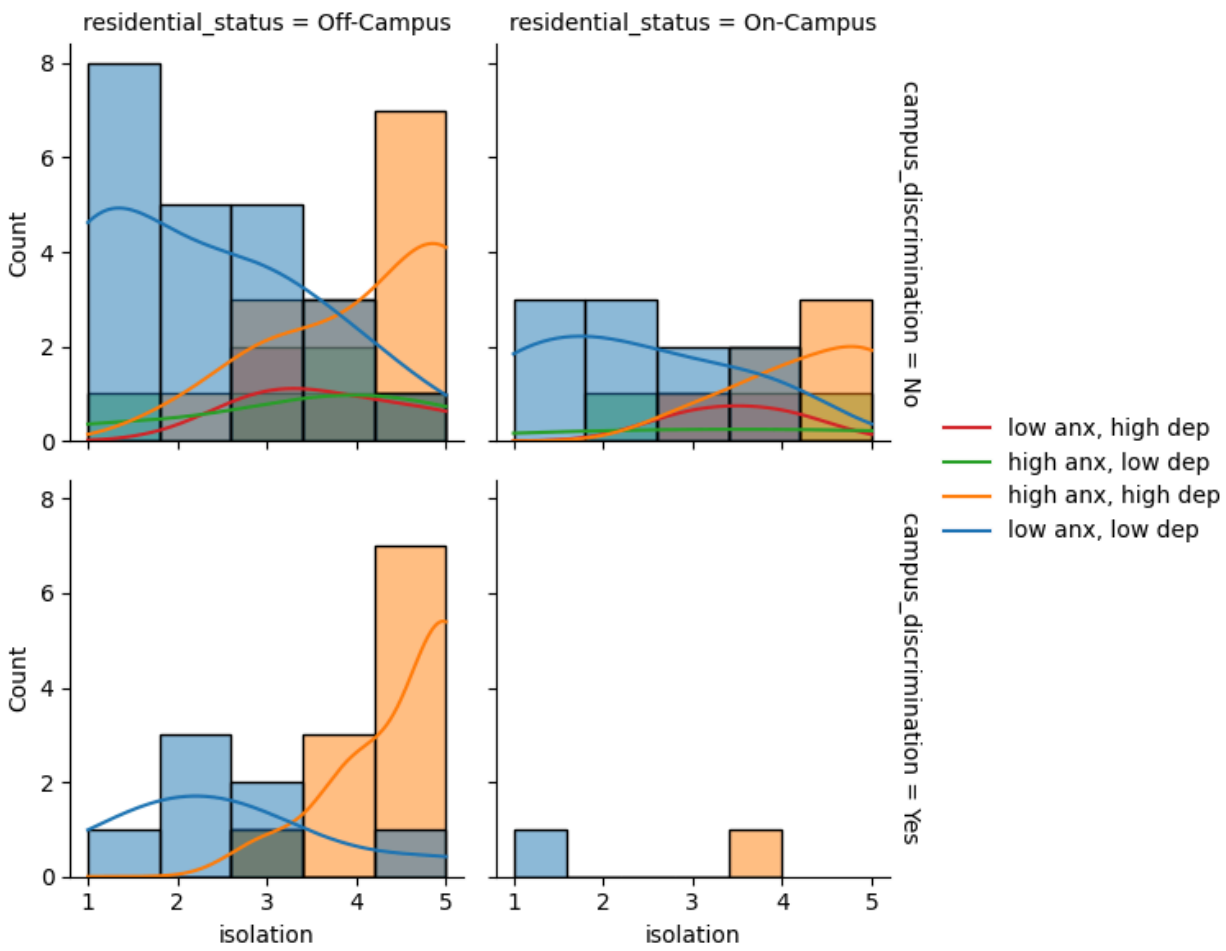

Figure 2) Histograms of Isolation, Plotted on Discrimination and Residential-Status Axis

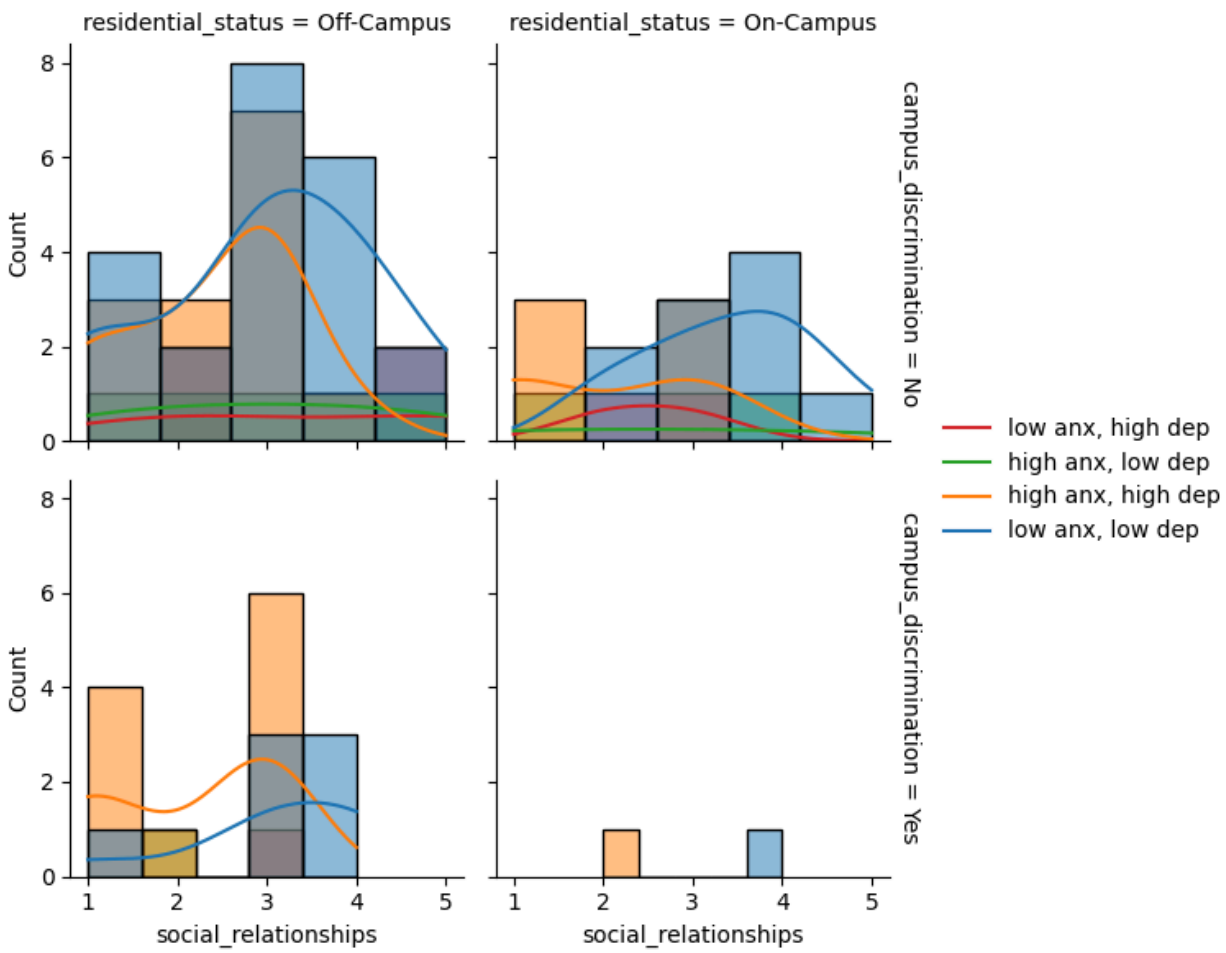

Figure 3) Histograms of Social Relationships, Plotted on Discrimination and Residential-Status Axis

Interestingly, despite the high negative correlation between social relationships and isolation seen on the heatmap, the relationships between social relationships and mental health rates did not seem as visually significant. Nonetheless, it can still be seen that most of those who have high anxiety and high depression seem to also have lower scores with the "social_relationships" variable, especially as compared to those with low anxiety and depression.

Before moving on to statistical tests, I also wanted to create some histograms to compare variables related to academic performance with rates of depression and anxiety, so I also created two histograms, one displaying the cumulative GPA (cGPA) of the participants, and one displaying the study satisfaction of the participants (the variable that, in the heatmap, appeared to have the highest correlation with higher rates of depression and anxiety). Contrary to the hypothesis, according to the histograms, students with high levels of depression and anxiety do not seem to necessarily have lower cGPAs. In fact, those who have high levels of depression and anxiety also appear to be the population with the highest number and percentage of students with cGPAs in the 3.5-4.0 range. Interestingly, the majority of students who have high levels of anxiety but low depression appear to have mostly cGPAs in the 2.5-3.0 range. Whether these differences are statistically significant is difficult to determine with graphical visualizations alone, so it is time to move on to statistical testing.

I used a few sets of two-sample T-tests to compare the averages in students' isolation, social relationship, study satisfaction, academic pressure, and future insecurity, between individuals with "high" (>3) and "low" (<3) rates of anxiety and depression. Two-sample T-tests were not something we covered in our class, but they are commonly used in psychology for comparing the means between two groups to find whether their differences are statistically significant. To do this, I did some research to find what libraries are usually used for statistical analysis tests, imported the libraries, and ran the tests. A snippet of the code can be seen below (Figure 4).

```
# Find whether there is statistical significance in the correlation between depression and anxiety;
# and isolation, social relationships, study satisfaction, academic pressure, and future insecurity.

# Separate the isolation data into four arrays, two containing high levels of depression and anxiety (>3)
# and two containing low (below average) levels of depression and anxiety (<=3)
high_dep_isol = survey_clean_data[survey_clean_data['depression'] > 3]['isolation']
low_dep_isol = survey_clean_data[survey_clean_data['depression'] <= 3]['isolation']
high_anx_isol = survey_clean_data[survey_clean_data['anxiety'] > 3]['isolation']
low_anx_isol = survey_clean_data[survey_clean_data['anxiety'] <= 3]['isolation']

t_stat_dep, p_val_dep = stats.ttest_ind(high_dep_isol, low_dep_isol)
t_stat_anx, p_val_anx = stats.ttest_ind(high_anx_isol, low_anx_isol)

print("T-statistic:", t_stat_dep, "(dep) and", t_stat_anx, "(anx)")
print("P-value:", p_val_dep, "(dep) and", p_val_anx, "(anx)")
```

```
T-statistic: 7.006156339014621 (dep) and 6.620260908598966 (anx)
P-value: 5.399858723338917e-10 (dep) and 3.060325866360363e-09 (anx)
```

Figure 4) Code snippet containing two-sample T-tests

A correlation between variables can be considered statistically significant when the p-value for a test is less than 0.05. For visual ease, I put the data into its own table, which is shown below (Figure 5).

	Variable	T-statistic (depression)	T-statistic (anxiety)	P-value (depression)	P-value (anxiety)
0	isolation	7.006156	6.620261	5.399859e-10	3.060326e-09
1	study_satisfaction	-1.737430	-1.718709	8.593341e-02	8.930763e-02
2	social_relationships	-2.755814	-3.340623	7.162231e-03	1.243402e-03
3	academic_pressure	3.120842	2.940193	2.463741e-03	4.223947e-03
4	future_insecurity	4.589371	4.156415	1.522933e-05	7.677760e-05

Figure 5) Table containing the T-statistic and P-values for variables of interest

With this table, we can see that there is a statistically significant difference between the level of isolation, social relationships, academic pressure, and future insecurity of students with high levels of anxiety and/or depression; however, there is not a statistical significance for study satisfaction.

IIIa) Discussion

The hypothesis that I came up with for this project was that Pakistani IT students with poorer mental health (evaluated through higher levels of depression and anxiety) are more likely to be socially uninvolved and do poorer academically. Our results showed that the part of the hypothesis related to social involvement is supported by data–individuals who had higher rates of anxiety and depression had significantly higher levels of isolation and lower scores in social relationships. However, results for academic success are less clear. While students with poor mental health do appear to show higher rates of academic pressure and insecurity about their future, it appears that they are still able to maintain relatively high cumulative GPAs and remain satisfied with their abilities to study. It is important to note, however, that correlation does not mean causation–whether poor mental health results in more pressure to function and succeed among mental hardships, or whether more pressure to succeed from external sources can influence student mental health remains unclear.

Nonetheless, the significant correlation of high isolation and low social involvement among students with poor mental health shows a need for some kinds of social interventions. It is possible that part of the reason why students may feel isolated when suffering from mental health problems is because of the stigma and lack of mental health literacy that is still prevalent in Pakistan. As such, this data presents a need for future research to look more closely at how encouraging the creation and maintenance of social relationships could help students feel less isolated, and how it can benefit their overall mental health.

IIIb) Limitations

While the dataset that I used was clean and manageable, this meant that I didn't really get the chance to use as many skills such as data cleaning and encoding as I would have liked. When doing my own research I know this will likely not be the case as many real-life surveys are often filled with missing values or outliers, especially if they are done on a larger scale or collected among the same group of participants over some period of time.

However, this dataset did have quite a few variables that could pose as risk factors for poorer student mental health, many of which I found sufficiently relevant to the main question (such as academic performance, social relationships, future insecurity, etc.). This is something I found to be lacking in other datasets that I was considering, which would commonly list students' problems with mental health (such as diagnoses), but not provide enough detail about other aspects of their life and well-being. Nonetheless, while most of the variables in this dataset were adequate for the purpose of this project, the lack of information on what the scales for other subjective but important variables such as "depression" actually meant made deriving any real conclusions from the data rather difficult. None of the surveys done using this dataset appear to use authentic measures such as Beck Depression Inventory, which is an example of a commonly used questionnaire for evaluating individuals' symptoms of depression. In the future, I'd like to find more specific and reliable survey measures when performing data analysis for a more realistic representation of real-life research.

V) References

Eisenberg, D., Hunt, J., & Speer, N. (2013). Mental health in American colleges and universities: Variation across student subgroups and across campuses. *The Journal of Nervous and Mental Disease, 201*(1), 60–67. https://doi.org/10.1097/NMD.0b013e31827ab077

How to find a p-value from a T-score in python?. GeeksforGeeks. (2024, January 22). https://www.geeksforgeeks.org/how-to-find-a-p-value-from-a-t-score-in-python/

Pollack, M. H. (2005). Comorbid anxiety and depression. *Journal of Clinical Psychiatry, 66*, 22.

Shafiq, S. (2020). Perceptions of Pakistani community towards their mental health problems: A Systematic Review. *Global Psychiatry, 3*(1), 28–50. https://doi.org/10.2478/gp-2020-0001

Zada, S., Wang, Y., Zada, M., & Gul, F. (2021). Effect of mental health problems on academic performance among university students in Pakistan. *International Journal of Mental Health Promotion, 23*(3), 395–408. https://doi.org/10.32604/IJMHP.2021.015903

VII) Acknowledgements

Special thanks to my friend Thomas, who isn't even in this class and did not contribute to this project at all but wanted a shoutout nonetheless.