How did universities respond to student mental health throughout COVID-19?*

Sidharth Gupta

03 May 2022

Abstract

Poor mental health is detrimental to learning capabilities, so it is extremely important that universities and colleges provide high-quality and accessible mental health resources. Using a survey dataset with over 10,000 responses from students across 8 universities, we evaluate the relationship between students struggling with their mental health, and how they perceive the quality of their schools mental health resources. We find that there is a weak association between these two features across all schools, and also make connections to how schools supported their students during COVID-19. Our work brings impactful results that university administration can consider to better support student health.

Keywords: mental health, data science, supervised learning, COVID-19

Contents

1 Introduction

A nation-wide U.S survey around college and university students found that almost half of those students experienced a psychiatric disorder in the past year; including alcohol, personality, anxiety, and mood disorders

*Code, simulation scripts, and data-sheet are available at: https://github.com/sidguptacode/Mental_Health_Analysis

(Blanco et al. 2008). Active Minds, a leading US-wide non-profit surveyed thousands of post-secondary students, and found that one third of students have felt depression to a degree where they struggled to function; three-quarters of students experienced a mental health issue; 80% described feeling overwhelmed; and 45% described a past feeling of hopelessness ("Covid-19: Student Survey" 2021). Despite these moving results, the Center for Collegiate Mental Health finds that only 25% of students experiencing mental health issues seek out resources, and at surveyed schools the counselor-to-student ratio on average ranges from 1:1000 to 1:3500 (n.d.).

Many studies in the literature (such as the ones listed above) highlight the severity of mental health issues in college and university. We conduct a data-driven analysis and build upon this literature in three points: first, we explore the causality of four variables related to student mental health; second, we investigate how these variables change throughout the COVID-19 pandemic; and third, we use these variables in a regression model to compare universities for the quality of their mental health resources. Our work fills the gap for understanding causes to the student mental health crisis, and brings insight to why some schools are handling the issue better than others.

We take a data-driven approach to this problem, using a survey dataset with over 10,000 responses from students across 8 universities. We analyze four key variables about mental health (shown in Table 1) using the R programming language (R Core Team 2021). To explore the variables, we group the dataset by university, plot the distribution of responses, and also plot features against each other in joint distribution as heatmaps. Our plots also group data from 2021 (when all the schools instructed online) and 2022 (when all the schools began instructing in-person). We then create a model for each university that helps to identify schools with more association on their mental health resources than others.

Our finding are important for university administration, as they directly impact academic participation and engagement for a large population of students. Several education research studies show that a student's poor mental health status is consistently related with their underperforming and underparticipation in academics (Markoulakis and Kirsh 2013). For example, depression is a significant predictor of lower GPA and higher probability of dropping out (Eisenberg, Golberstein, and Hunt 2009). Our data-driven analysis visualizes these complex mental health variables that are often seen in the literature, and makes them more understandable in a student context. In addition, to the best of our knowledge, there does not exist a benchmark that compares universities for the quality of their mental health resources. Our investigation is a first step towards creating such a benchmark. Such a comparison is impactful, as university staff can look at other schools and take in new ideas on how to better support their students.

Our paper is organized as follows. In the data section, we rigorously introduce the mentioned mental health dataset. We describe how data is collected, what the data looks like, the features that will be investigated, current literature in mental health science, biases present in the dataset, limitations, and motivations behind studying this dataset despite its flaws. In the model section, we propose possible causal relationships between the analyzed dataset features, introducing directed-acyclic-graphs (DAGs) to help reason about these causal relationships. We investigate these DAGs using both a linear regression and random forest model, and explain how factors such as multicollinearity affect these models. In the results section, we interpret all of our proposed models with respect to the DAGs they probe, and thoroughly visualize the dataset. We group the dataset by universities so that we can compare schools, and investigate the aforementioned DAGs with heatmaps of joint distributions. In the discussion section we describe implications of these results with respect to current literature in mental health science, and future analyses that can extend our work in investigating causes behind decreased student mental health.

2 Data

We utilize a dataset that was collected by a voluntary student-run project called the Aphrodite project. The Aphrodite project provides a service where students can sign up, fill in a personality survey, and get matched with another student for a platonic or romantic relationship. The Aphrodite project has been running annually at 7 universities in North America and at 1 university in Asia: Harvard University, HKU (Hong Kong University), McMaster University, Massachusetts Institute of Technology (MIT), University

Table 1: One record from optional mental health section. Each column is rated on a scale from 1 (strongly disagree) to 7 (strongly agree)

I am currently struggling with my mental health	Students can access mental health resources and support at my university easily	I have high self esteem	I'm facing problems in my relationships due to COVID-19
4	4	7	4

Table 2: One record from the demographics section.

Participant gender identity	Participant year of study	Participant university	Year data was collected
Cisgender Woman	Master's	Harvard	2021

of British Colombia (UBC), University of California Los Angeles (UCLA), University of Toronto (UofT), and the University of Waterloo (UWaterloo). In addition to personality questions, at the end of the survey participants are asked a set of optional questions related to their mental health. It is made clear in the survey that the mental health questions are purely optional and have no impact on the matching result that the service provides. The first row of these optional mental health questions are shown in Table 1. Participant demographics are also collected, as visualized in Table 2. Histograms that visualize the responses for these questions are shown in Figure 1. There are additional questions asked in the Aphrodite dataset, but in this analysis only kept those questions prevalent in the mental health literature.

These specific mental health questions were chosen for their prevalence in the literature, as well as their unique timing being collected during the COVID-19 pandemic. The question about relationship problems from COVID-19 stems from studies around COVID isolation. Multiple studies have shown that spending time in social relations and engaging in talk therapy are essential components of maintaining mental health (Gilburt, Rose, and Slade 2008). One of the many unfortunate consequences from the COVID-19 pandemic has been its affect in removing these components and bringing social isolation to many. Indeed, a survey from the World Health Organization found that 29% of respondents reported moderate to severe anxiety symptoms due to isolation from COVID-19. (Matias, Dominski, and Marks 2020)

Additionally, self esteem is a highly associated factor of mental health. For university students in particular, low self-esteem can be caused due to the imposter phenomenom, differences in socioeconomic statuses, and academic performance (Sonnak and Towell 2001). These reasons motivate the question about self-esteem. The remaining two questions ask if the student is struggling with their mental health, and how the student perceives the quality of their school's mental health resources. These last two questions play an important role as target variables in our analysis.

Another possible dataset to support this analysis would be to analyze the text and sentiment of online university forums, such as reddit. We discuss possible extensions with these forums in the Discussion section, but omit their analysis in this study due to the many nuances and complexities involved with investigating a dataset of that scale.

There are many possible data biases with this collected data. First, the data is collected as a part of a student matchmaking service, so only students who are comfortable engaging in such a service would have participated. Because of this fact, it is possible that certain demographics of students are entirely excluded from the dataset. For instance, students who are more introverted and prefer not to meet strangers would not participate in the service. Additionally, the mental health questions are asked in the context of matchmaking, which can introduce some unconscious bias. The service does state that the mental health questions are

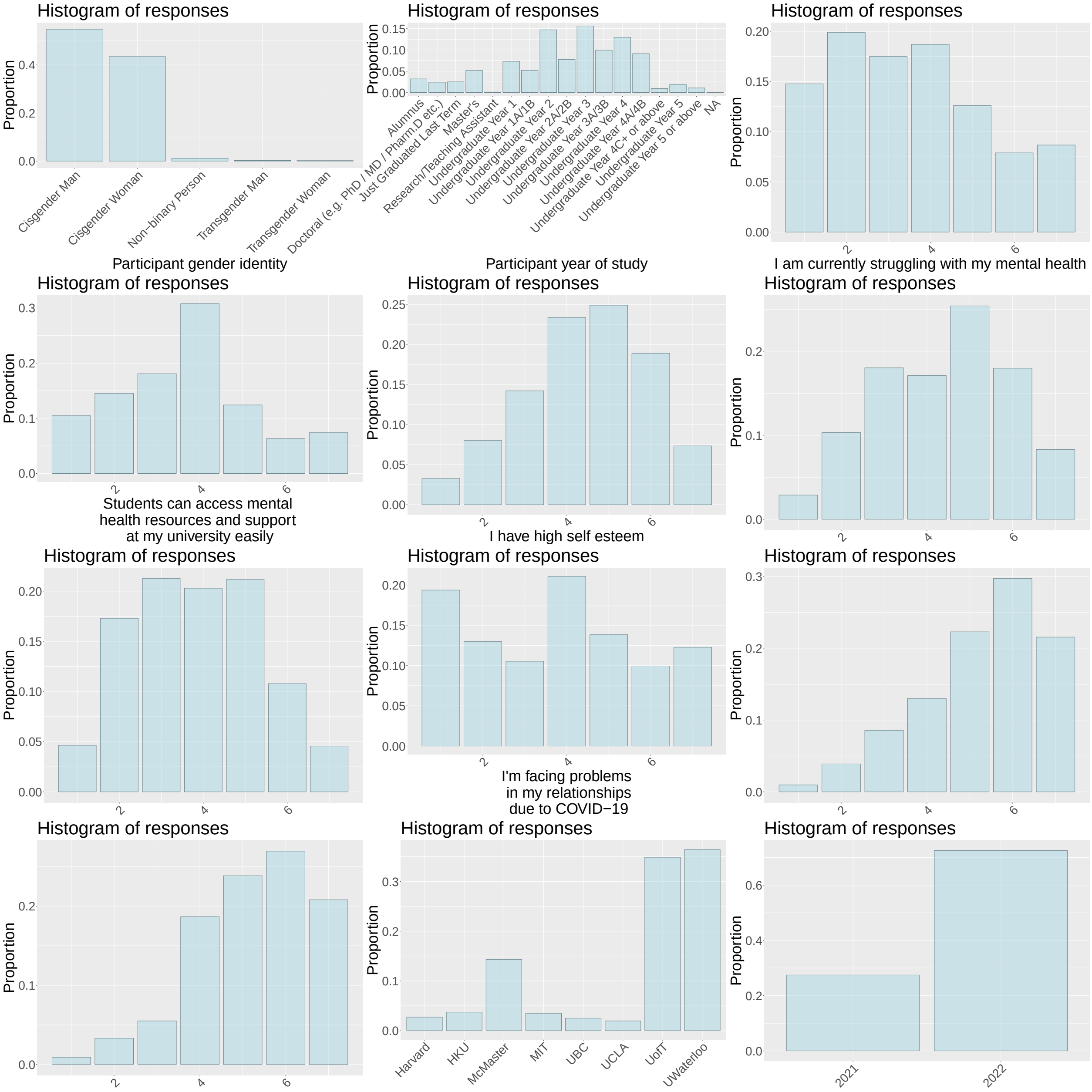


Figure 1: Histograms for each column in the considered Aphrodite dataset

optional and have zero impact on the matching result; however the context of matching may push participants to answer less critically to these sensitive questions. Finally, our results compare universities for the quality of their mental health resources, however as our histograms in Figure 1 show, some schools have a greater number of responses than others. As a result, for schools such as UCLA, UBC, and Harvard, the true distribution of student mental health experiences is likely not captured.

Despite these drawbacks, there are multiple rewards to be had in analyzing this dataset, as long as the analysis is careful about the mentioned biases. First, to the best of our knowledge it is the only dataset that spans universities from Canada, the U.S, and Hong Kong, and collects the same survey data from each of those schools. Second, the results from our analysis, although hindered from bias, do converge with studies in the mental health literature. And third, the positive results from this analysis can act as a gateway for more resources to be allocated for data collection of student mental health resources at a large scale. We outline possible extensions and incentives for future data collection efforts on student mental health in the Discussion section.

For our implementation, we use the R programming language (R Core Team 2021), and perform the following processing steps. We load the data with the R Tidyverse and tidyr libraries (Wickham et al. 2019), (Wickham 2021b), and clean the table headers using the R janitor library (Firke 2021), the dplyr library (Wickham et al. 2021), and the forcats libary (Wickham 2021a). We construct our model summary tables using the R modelsummary library (Arel-Bundock 2022), and developed our random forest model using the randomForest library (Liaw and Wiener 2002). The ggplot library (Wickham 2016) was used to generate all of our figures.

3 Model

3.1 Causality hypothesis

The relationships between mental health, self esteem, relationship problems from COVID-19, and perception of university mental health resources are very complex. We hypothesize several such causal relationships, and probe them with regression models and visualizations. To help reason about these causal relationships, we use directed-acyclic-graphs (DAGs). DAGs represent dataset features as nodes, and connect those nodes with uni-directional arrows such that there are no cycles (that is, it is not possible to start at a node, follow a path of outgoing arrows, and arrive at the starting node). Each unidirectional arrow represents a causal relationship. For example, in the DAG, let $x_{sEsteem}$, $x_{mHealth}$, x_{covid}, and $x_{uniSupport}$ each represent the dataset features: “I have high self esteem,” “I am currently struggling with my mental health,” “I’m facing problems in my relationships due to COVID-19,” and “Students can access mental health resources and support at my university easily.”

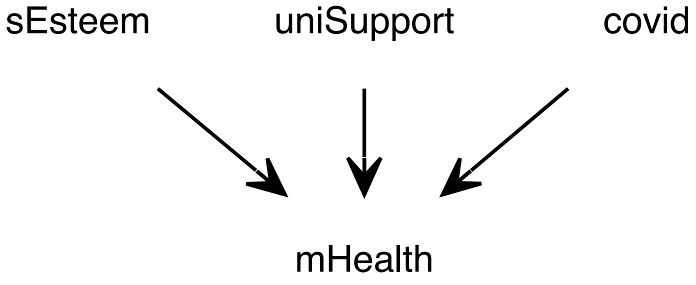


Figure 2: One possible directed acyclic graph representing the causality relationship between the dataset features

We can see that the DAG shows $x_{mHealth}$ to be dependent on $x_{sEsteem}$, x_{covid}, and $x_{uniSupport}$. We hypothesize the dependency on $x_{sEsteem}$, x_{covid} based off of the discussion in the Data section. Additionally, we hypothesize a dependency on $x_{uniSupport}$, because if a student believes their school’s mental health resources are not good, then they may be less likely to use those resources to improve their health. If a student believes their school’s resources are good, they may be more likely to use them, which would reduce the degree of $x_{mHealth}$.

3.2 Linear regression

We can formulate a linear regression model to represent this DAG. Because each feature is rated on an ordered-likert scale, using a regression model over a classification model makes sense in this context. Let $\hat{y}$ represent a prediction of $x_{mHealth}$ based off the mentioned variables. Let y represent the true value of $x_{mHealth}$. Define $\hat{y}$ as:

$$\hat{y} = w_0 + w_1 \cdot x_{sEsteem} + w_2 \cdot x_{covid} + w_3 \cdot x_{uniSupport}$$

To implement this linear model, we use the stats package built in R (R Core Team 2021). This package finds the set of weights w_0, w_1, w_2, w_3 such that the following expression is minimized:

$$\sum_{i=0}^{N} (\hat{y}_i - y_i)^2$$

, where N represents the size of the dataset, y_i is the true value of y in the $i'th$ record in the dataset, and $\hat{y}_i$ is the predicted value of y using the $i'th$ record of the dataset.

Once the R stats package (R Core Team 2021) finds such a minimal set of weights that define this linear model, we can interpret them to see how well the DAG holds up. We group the dataset by university, and define a model for school's data. The resulting model summary statistics are shown in Table 3.

Table 3 shows optimized values for w_0, w_1, w_2, w_3 as Intercept, self_esteem, covid_relationships, and access_resources, respectively. An R^2 value is included for each school, which is each model's coefficient of determination. We do not use a train-test set because some schools have less samples – instead, we evaluate the fit of each model using this R^2 value. This R^2 value can give insight on how well the target variable is correlated with the input variables. Typically in the literature, we say that R^2 values between 0.2 and 0.4 show low to moderate correlation, and that is what we see for all of the models shown above, thus providing support for our DAG.

Each weight is also assigned a value p, which we denote as the p-value. The p-value represents the probability that a null hypothesis is correct – that is, the probability that the feature associated with weight w_i has an insignificant impact on the model prediction result. If the p-value is low, we can reject this null hypothesis, and affirm that the feature is indeed significant for model predictions. A standard threshold for calling a p-value "too-low" is 5%. Our model results and p-values are shown in the Results section, as Table 3 and 4.

3.3 Random forests

There are some shortcomings with the DAG, and the primary flaw is that self esteem, perception of university resources, and relationship issues from COVID are treated as independent, when certainly there is a complex dependency among these attributes. Indeed, our visualizations in the Results section do show a joint relationship when these features are plotted against each other in a heatmap. This multicollinearity can introduce numerical irregularities as we are computing p-values for the weights w_0, w_1, w_2, and can make the meaning of these p-values less robust.

We address these issues by also modelling the data with a random forest model, that is more resilient to issues in multicollinearity. To explain random forests at a high level, we will start by explaining decision trees. At a high level, decision trees are models that ask a binary question about each feature, such as "is this feature x greater than or less than some threshold α?" These questions split the dataset into halves, and are repeatedly asked until each record in the dataset is split into a group. These groups then represent predicted labels for each point in the dataset. The order of the questions and the values of α are learned in an algorithm such that they maximize an information gain function. A random forest ensembles multiple decision trees, and does regression by averaging each decision tree's prediction. Random forests ensure that each decision tree used is sufficiently different, by training each with a different bootstrapped version of the original dataset. The concern of multicollinearity is less prevalent with random forests, because thousands of decision trees are used, and the probability of many models exploiting the same correlations in the input (while at the same time each being sufficiently different) is certainly lower than one model exploiting those correlations.

Same as in linear regression, we group the dataset by university, and train a random forst model for each school. We do not use a train-test split to evaluate model performance because some schools have limited datapoints, and instead use MSE as a proxy for fit. MSE is defined as the averaged mean squared error between the prediction and the label, across all predictions, as in linear regression. After training, we interpret each school's model with respect to the DAG. For each trained random forest, we display it's accuracy, and a measurement %IncMSE for each input feature. %IncMSE measures the percent increase in mean-squared-error after the $i'th$ feature gets randomly permuted (i.e, randomly shuffled). The higher the value of this attribute, the more significant the feature is for the prediction task.

4 Results

Table 3: Linear regression models for predicting students struggling with their mental health.

	Harvard	HKU	McMaster	MIT	UBC
(Intercept)	3.745	2.553	4.804	4.282	4.148
	p = 0.000	p = 0.000	p = 0.000	p = 0.000	p = 0.000
self_esteem	−0.140	−0.276	−0.438	−0.326	−0.328
	p = 0.031	p = 0.000	p = 0.000	p = 0.000	p = 0.000
covid_relationships	0.321	0.363	0.322	0.322	0.378
	p = 0.000	p = 0.000	p = 0.000	p = 0.000	p = 0.000
access_resources	−0.270	0.082	−0.097	−0.171	−0.124
	p = 0.000	p = 0.155	p = 0.000	p = 0.001	p = 0.070
Num.Obs.	278	381	1466	358	257
R2	0.263	0.190	0.306	0.272	0.248

Table 4: Linear regression models for predicting students struggling with their mental health.

	UCLA	UofT	UWaterloo
(Intercept)	6.193	4.556	4.601
	p = 0.000	p = 0.000	p = 0.000
self_esteem	−0.505	−0.425	−0.395
	p = 0.000	p = 0.000	p = 0.000
covid_relationships	0.159	0.306	0.282
	p = 0.004	p = 0.000	p = 0.000
access_resources	−0.215	−0.068	−0.112
	p = 0.001	p = 0.000	p = 0.000
Num.Obs.	199	3570	3730
R2	0.273	0.263	0.239

In this section, we will display and interpret the aforementioned models, as well as visualize correlated features in the dataset. To start, we visualize the linear regression models for each school, shown in 3 and 4. In our model results, we can see that all features have a p-value less than 5%, except for $x_{uniSupport}$ at HKU and UBC (that feature is represented by access_resources). Our results from the random forest models are shown in 3. The average MSE for each school's model is between 1.9 and 2.5, meaning that the random forest on average makes predictions that are between 1.3 and 1.6 Likert points off of the target.

Next we perform a visualization of the dataset features, and explore their correlated properties. We do so by plotting the joint distribution of responses for $x_{mHealth}$ and $x_{uniSupport}$. These features are chosen as they have shown significance in our previous analyses. Interactions between these features and $x_{sEsteem}$ are shown in the Appendix section.

5 Discussion

5.1 Comparing schools requires more analysis

In our analyses we do notice some differences in the grouped universities, One interesting finding comes from the high p-value for $x_{uniSupport}$ at HKU and UBC in linear regression, as shown in Table 3. The feature $x_{uniSupport}$ measures a student's perception on the quality of their school's mental health resources. The p-values claim that this feature may be independent, or have a very weak dependency on student mental health scores in these schools. The random forest results further probe the role $x_{uniSupport}$ at different schools. The trained random forest models, visualized in Figure 3, are trained to predict between 1.3 and

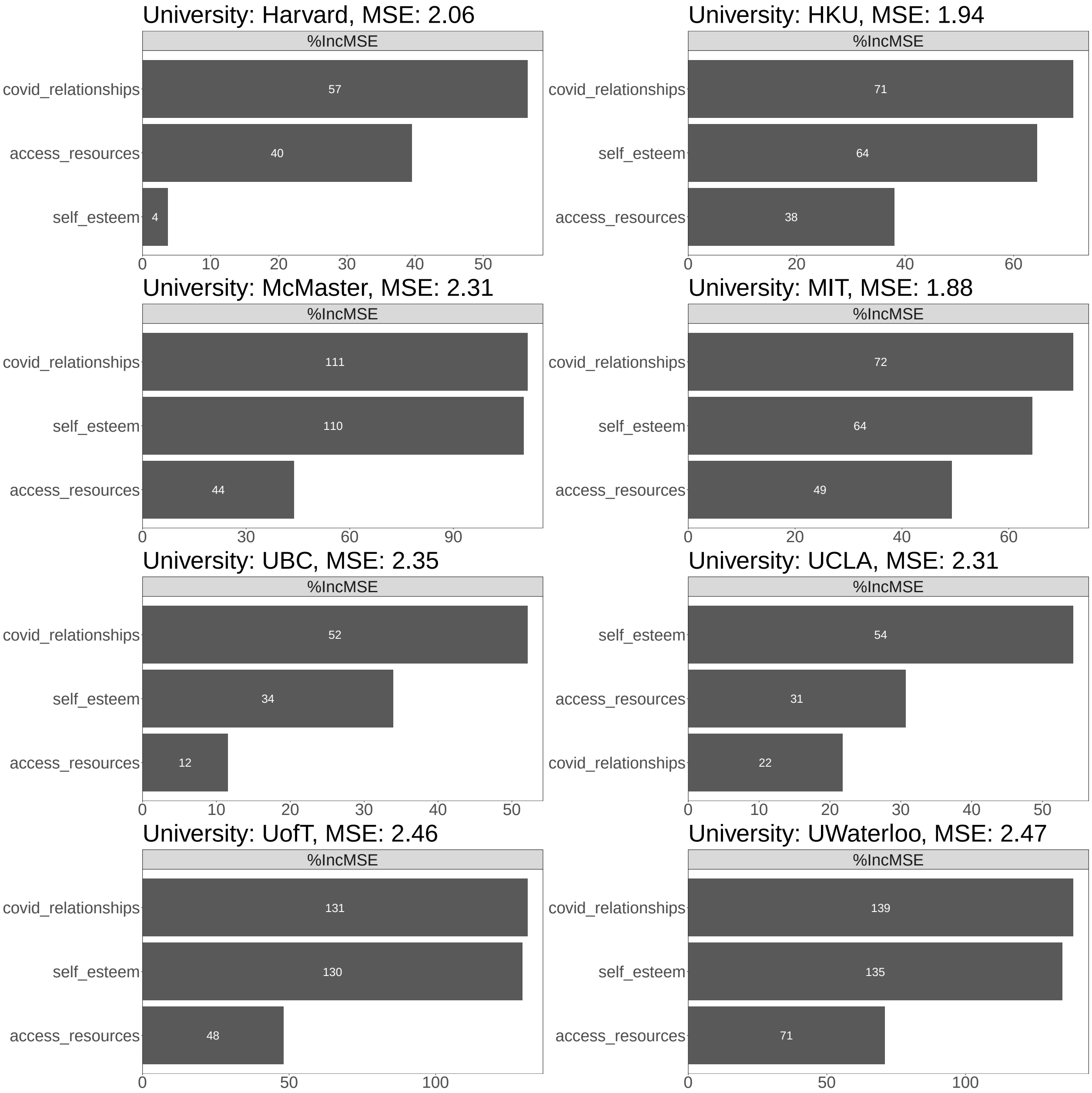


Figure 3: Random forest % increase in MSE plots, grouped by university, plotting average MSE

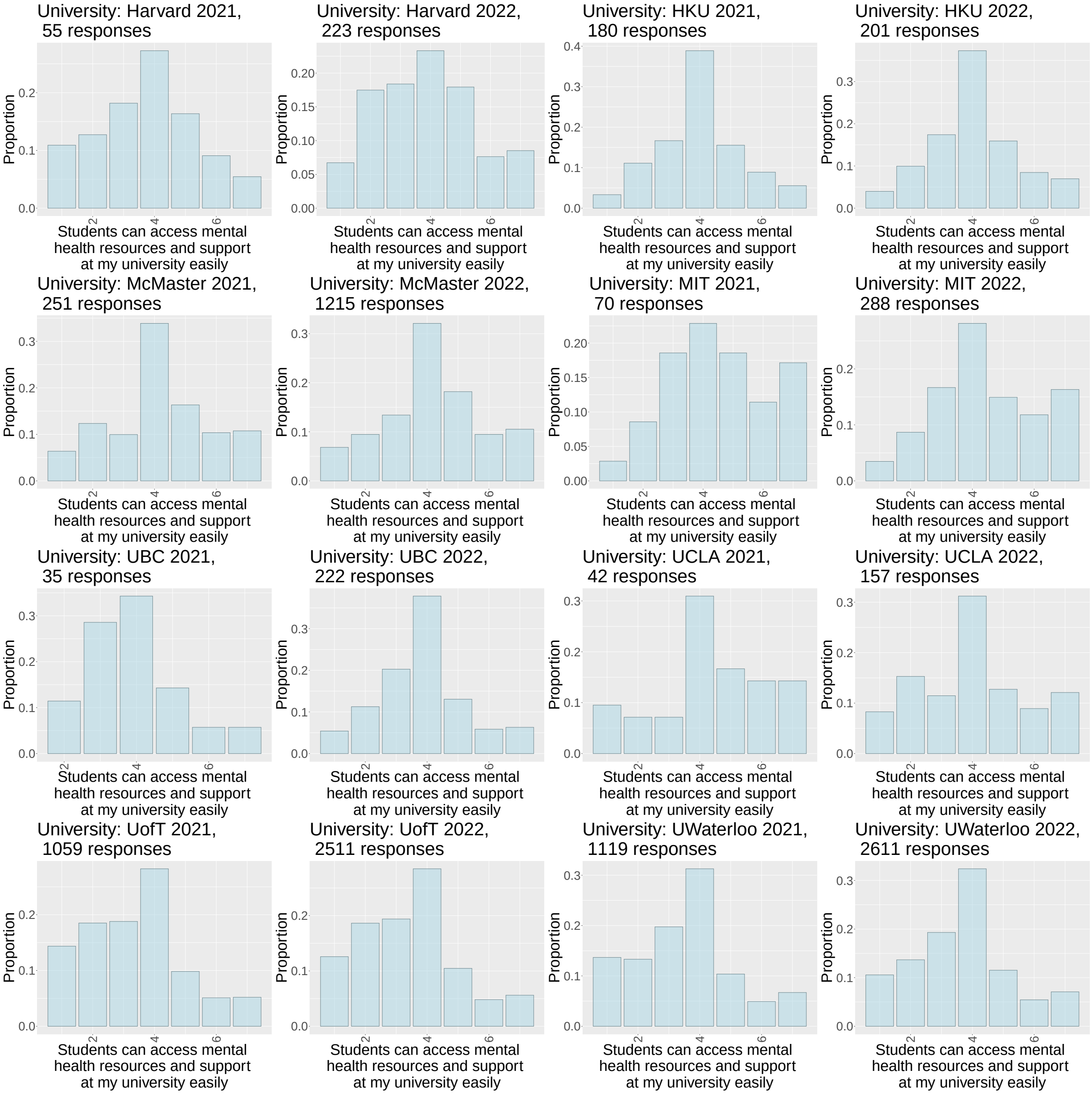


Figure 4: Distribution of agreement with mental health resources accessibility

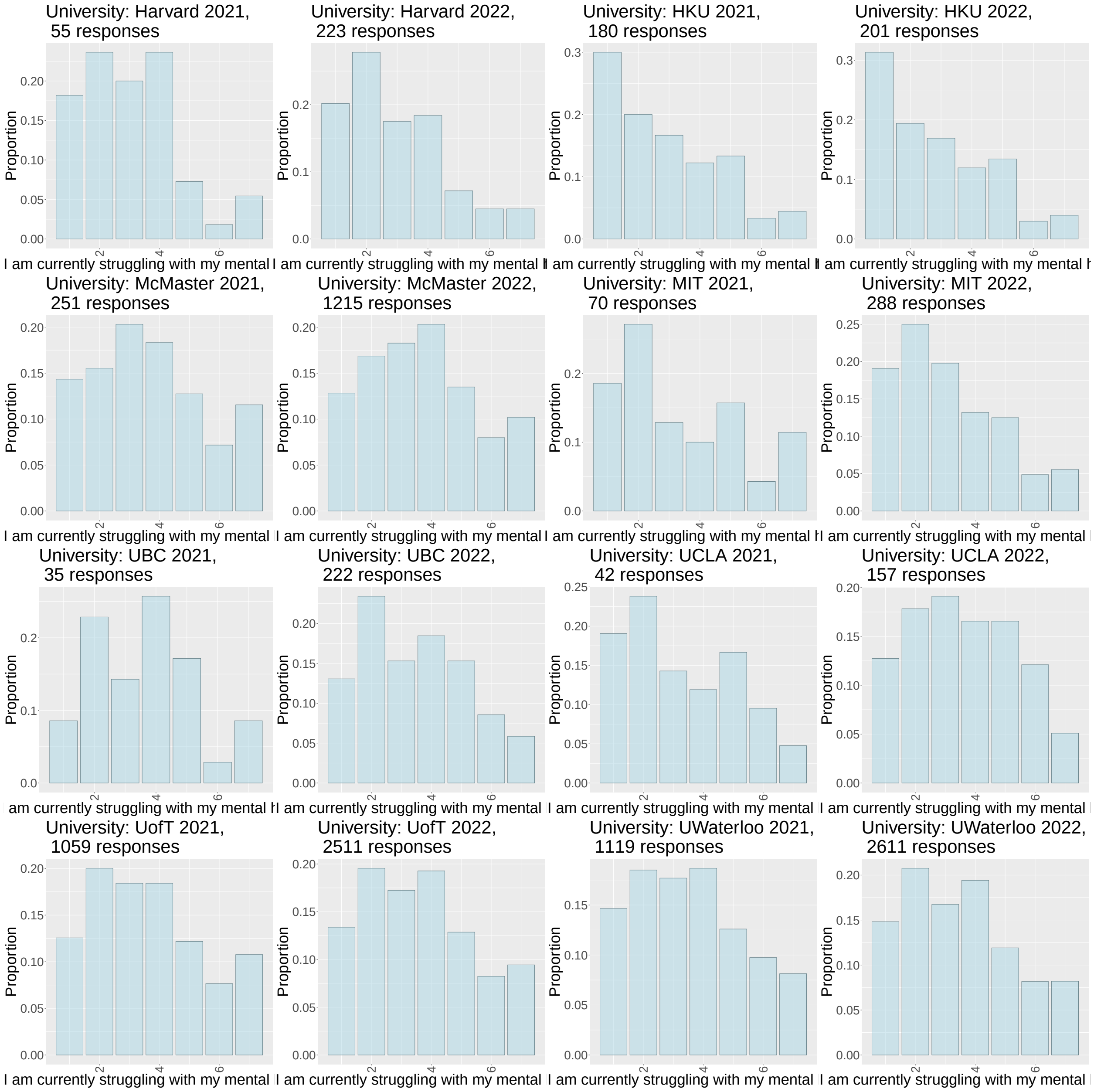


Figure 5: Distribution of agreement with struggling student mental health

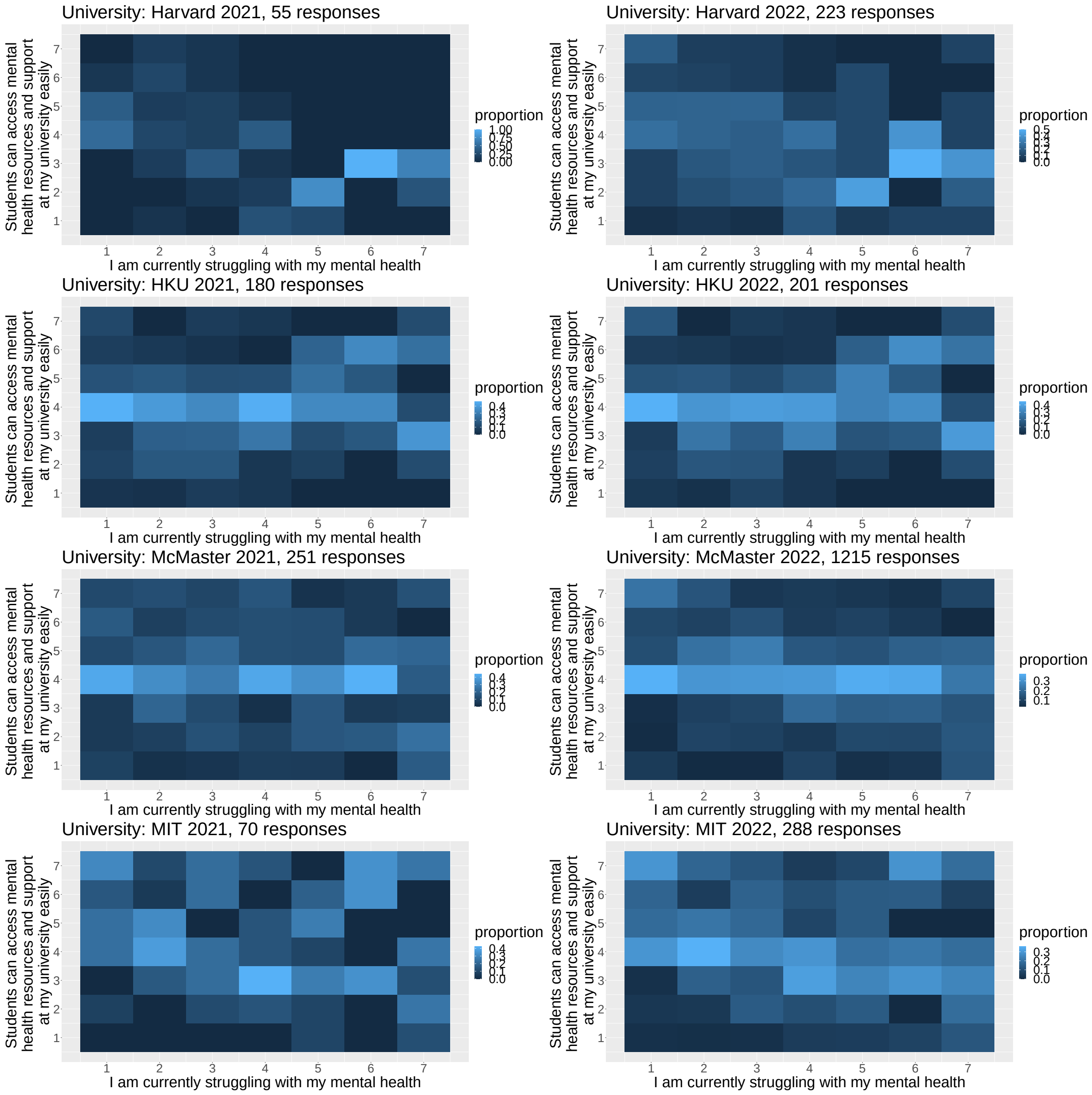


Figure 6: Heatmap of struggling mental health score vs. resource accessibility

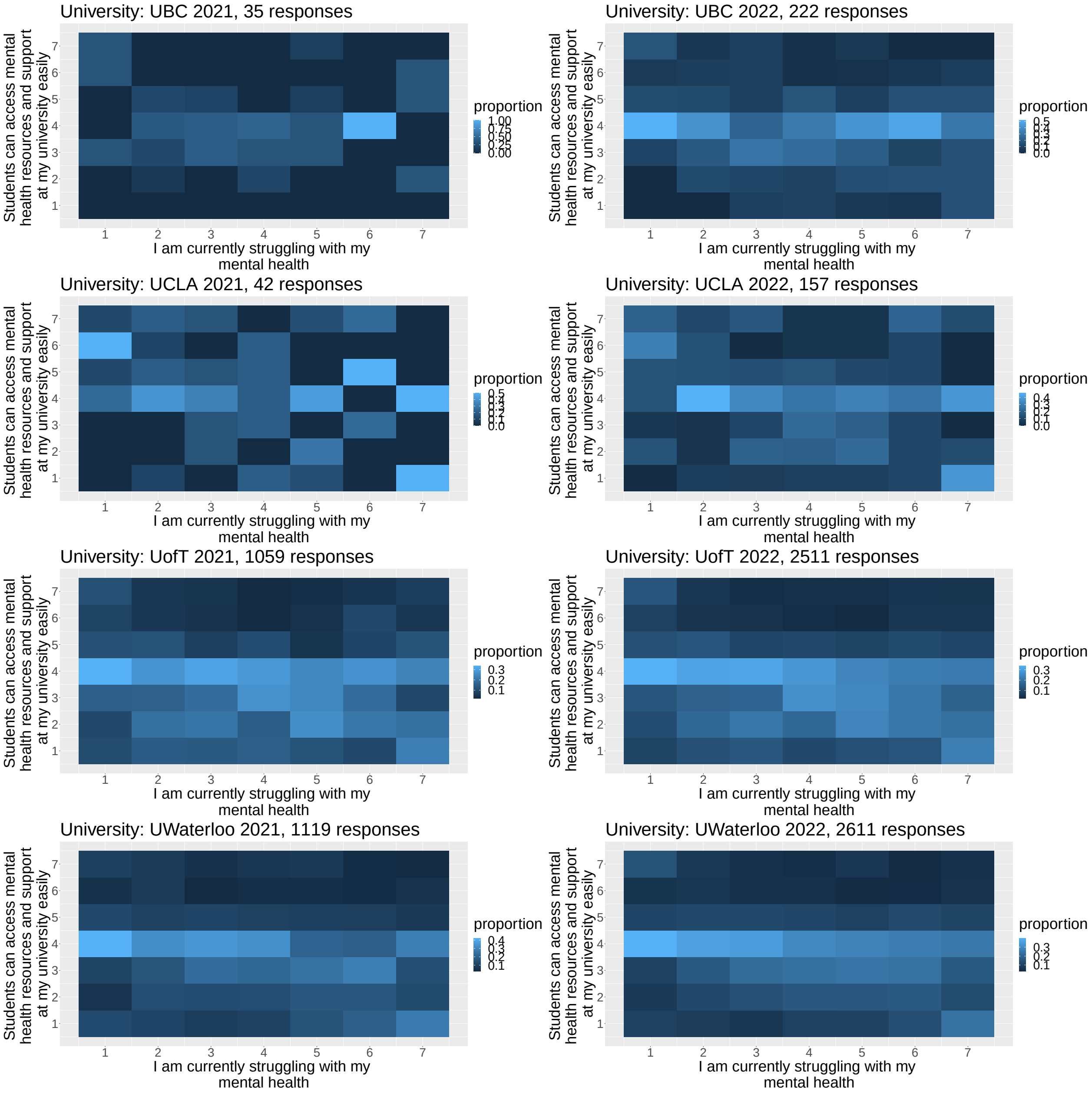
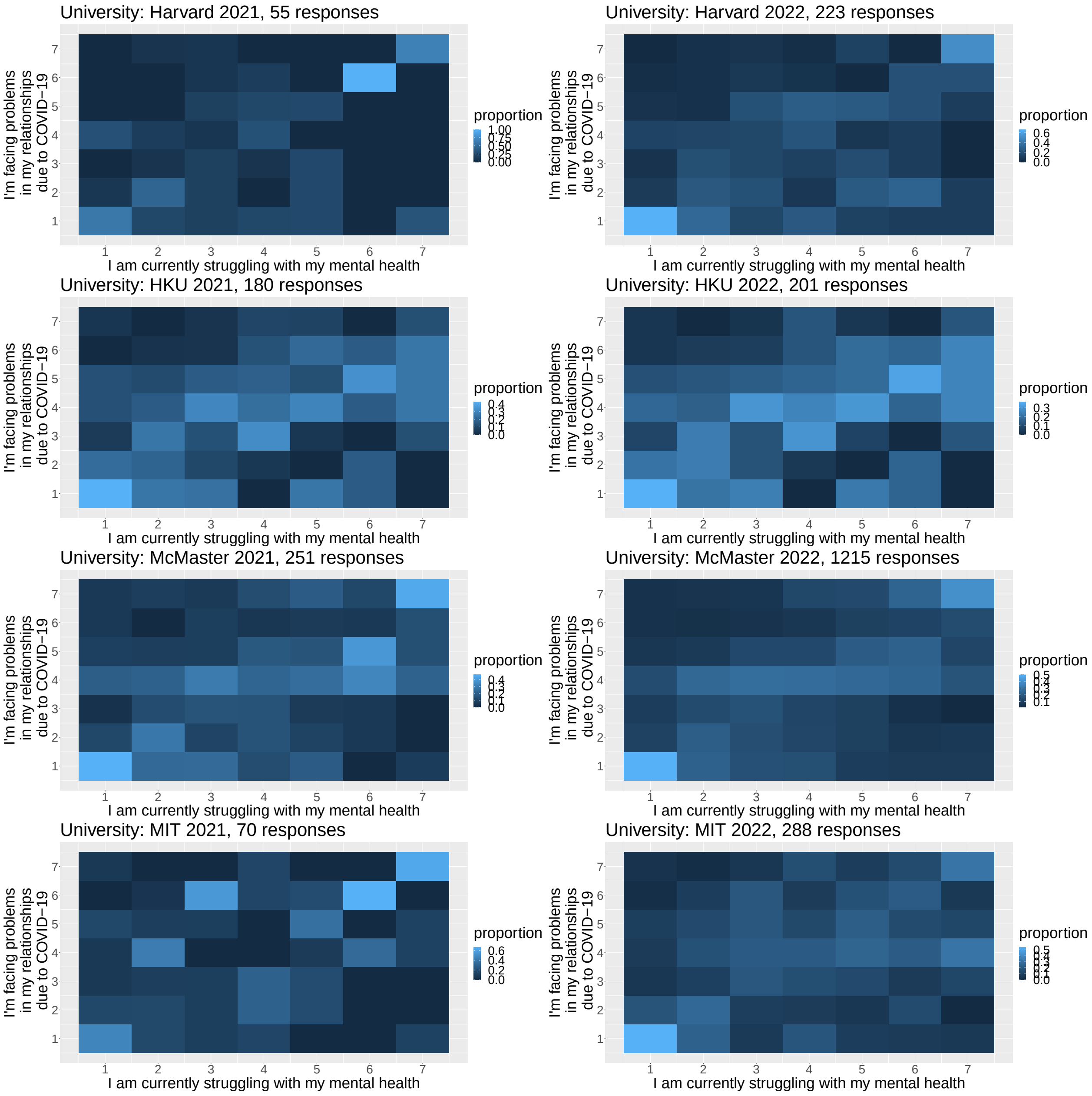


Figure 7: Heatmap of struggling mental health score vs. resource accessibility



University: Harvard 2021, 55 responses

University: Harvard 2022, 223 responses

University: HKU 2021, 180 responses

University: HKU 2022, 201 responses

University: McMaster 2021, 251 responses

University: McMaster 2022, 1215 responses

University: MIT 2021, 70 responses

University: MIT 2022, 288 responses

Figure 8: Heatmap of struggling mental health score vs. COVID-19 relationship impact

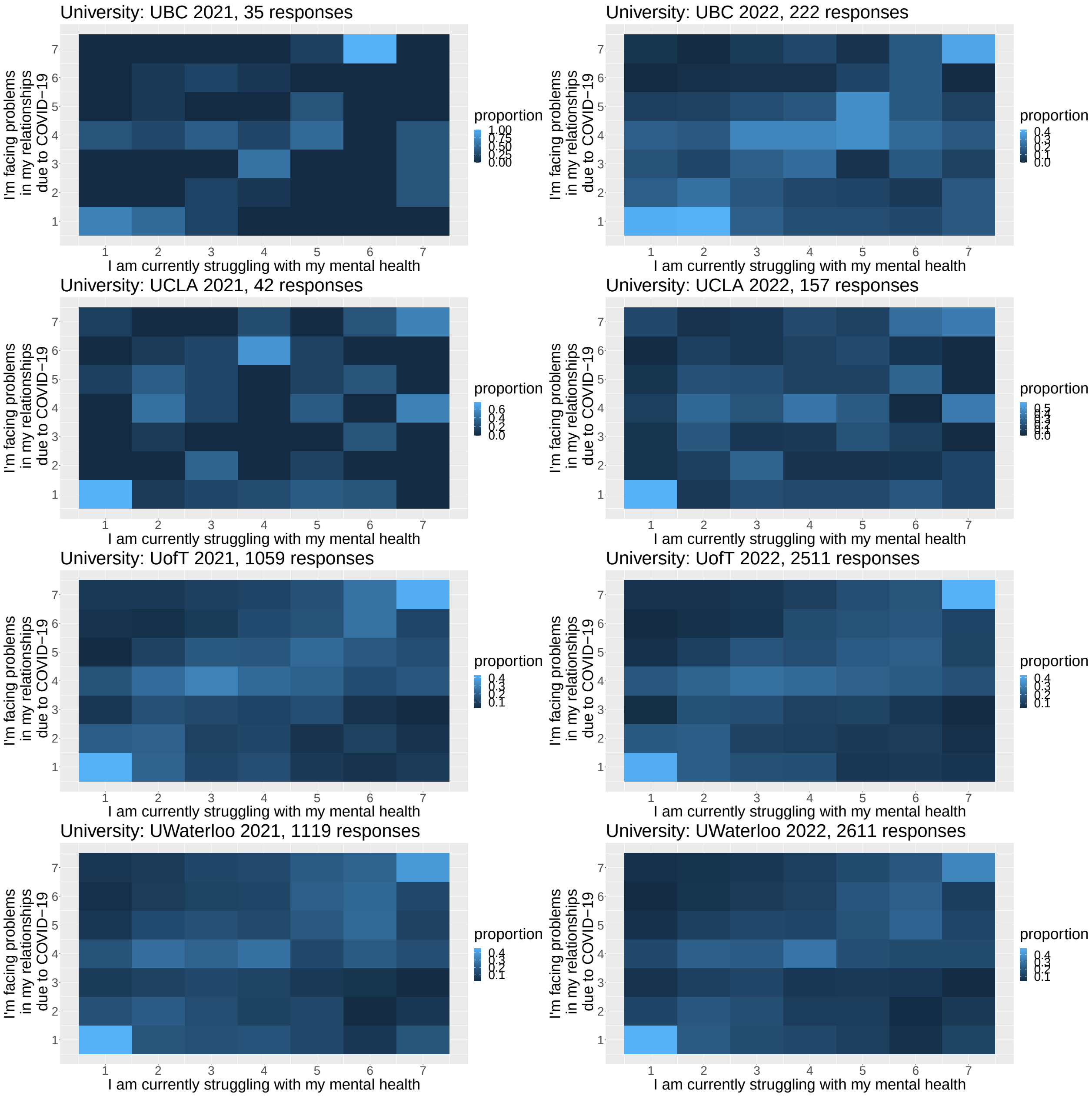


Figure 9: Heatmap of struggling mental health score vs. COVID-19 relationship impact

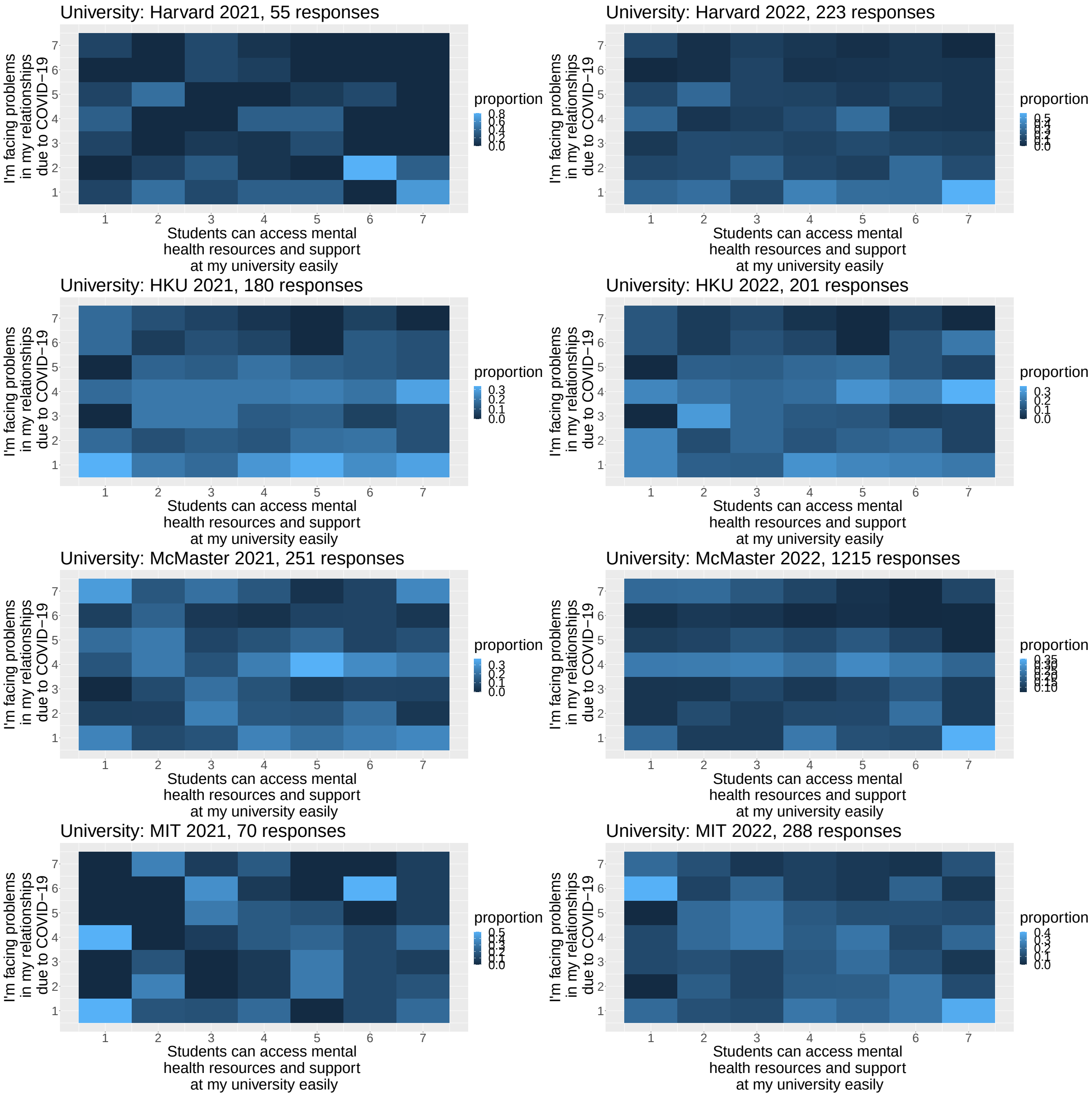


Figure 10: Heatmap of resource accessibility vs. COVID-19 relationship impact

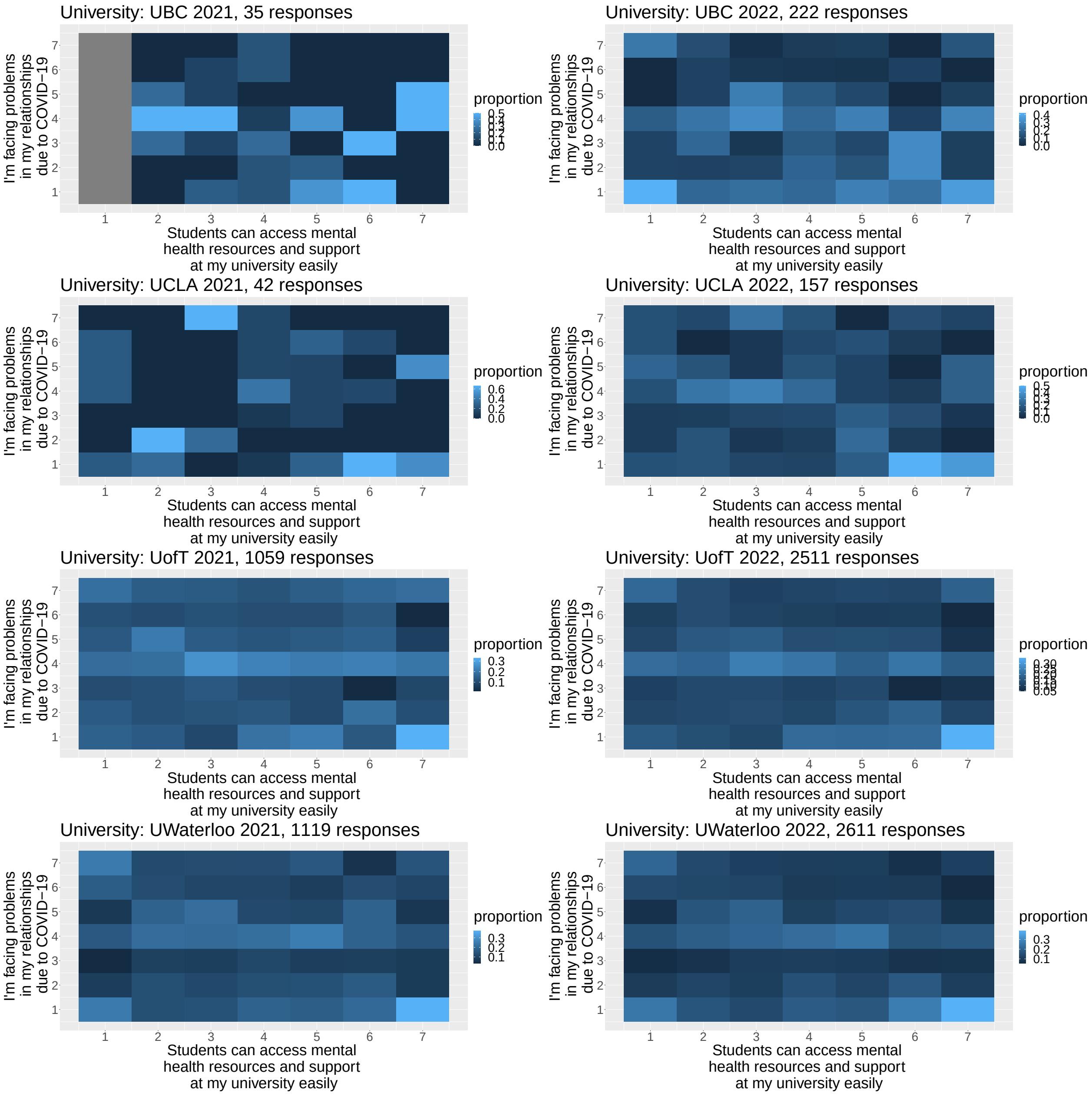


Figure 11: Heatmap of resource accessibility vs. COVID-19 relationship impact

1.6 Likert points off of the target. In the context of this dataset, that kind of prediction power would be able to identify students with positive and negative mental health relatively well, but would not perform as well in granular situations, such as identifying a student with poor and very poor mental health. The plotted %IncMSE for each feature is shown in Figure 3, and supplements the linear regression results very well with new information. With these plots, we can see more clearly that the although the role of $x_{uniSupport}$ (represented by access_resources) is significant in linear regression, it is certainly not as significant as $x_{sEsteem}$ or x_{covid} for most schools. Additionally, UBC and HKU have %IncMSE values for $x_{uniSupport}$ of 12% and 38%, meaning that permuting that feature increases the average MSE by 1.12 and 1.38 the original amounts, respectively. We can multiply the original MSE by these measurements, and find that the UBC model would predict 0.3 extra points away from the target if this feature is permuted, and the HKU model would predict 0.76 extra points away from the target. Both of these quantities are still small, and further support possible evidence for this feature having weaker dependency at UBC and HKU. However our dataset limitations and biases certainly could be playing a role in us seeing these results, so no concrete conclusions can be reliably made. There are only 381 and 257 responses collected from HKU and UBC, respectively, so a greater sample size would allow us to gain more insight about this finding.

5.2 Students perceive their schools to be doing better than they actually are

Our analyses ultimately show that the feature $x_{uniSupport}$ has a weak relationship with student mental health scores. For every school, the random forest plots show this feature to have the lowest value of importance in the model. Even across models, the importance metric are low for this, and if this feature was permuted with random values, the model would still have a reasonable amount of predictive power left. The results from the model are further emphasized when we plot the joint distribution of this feature with other features. In Figures 6 and 7, we can see that ratings for this feature do not significantly differ as student mental health scores decrease. In fact, for schools such as UofT, UWaterloo, McMaster, and HKU, the scores for $x_{uniSupport}$ center around the neutral value of "4" for all mental health scores, with a small amount of variance. This can imply that most respondents don't know what the quality of their university's mental health resources are like, even if they have poor mental health. A reason for students not having this information can potentially be caused by a lack of notice or awareness given by the schools.

5.3 Relationship problems due to COVID are common, and don't diminish when school transition to in-person

When we look at the extremes of students who have a very low or very high mental health score, we can see in the data visualizations that relationship problems due to COVID-19 are a highly associated feature. This behavior is shown clearly in Figures 8, and 9. The role of this feature is shown to have a significant impact on the models, however its significance is perhaps driven by these extreme cases. What is most interesting, is that the aforementioned plots do not change much from 2021 to 2022. This can imply that even though universities are switching from online to in-person, relationship problems from COVID-19 still persist. Additionally, there is little relationship between relationship problems and the feature $x_{uniSupport}$, which can potentially imply that universities are not providing enough resources to deal with the social isolation caused by COVID-19.

5.4 Extensions with analyzing university subreddits

Reddit is an online platform that hosts forums for different communities. On reddit there are forums for many universities around the world, that each store thousands posts every month. The scale of the data on university subreddits is very large, and there is certainly information to be found about student mental health. However, data on reddit is in the form of unstructured text and posts, and spans all topics related to college and university life – not just mental health. As a result, analyzing the reddit text dataset would require a large amount of resources to handle the noisy environment. Still, information about how students are affected by COVID-19 and struggles that students experience all exist on reddit, and as such exploring such a dataset could help address some of the flaws in this study. A simple approach could be to use natural language processing techniques to gauge sentiment analysis on posts at different university subreddits, and comparing schools using those sentiment scores. Those sentiment scores could also be treated as features for the analysis shown in this study.

6 Appendix

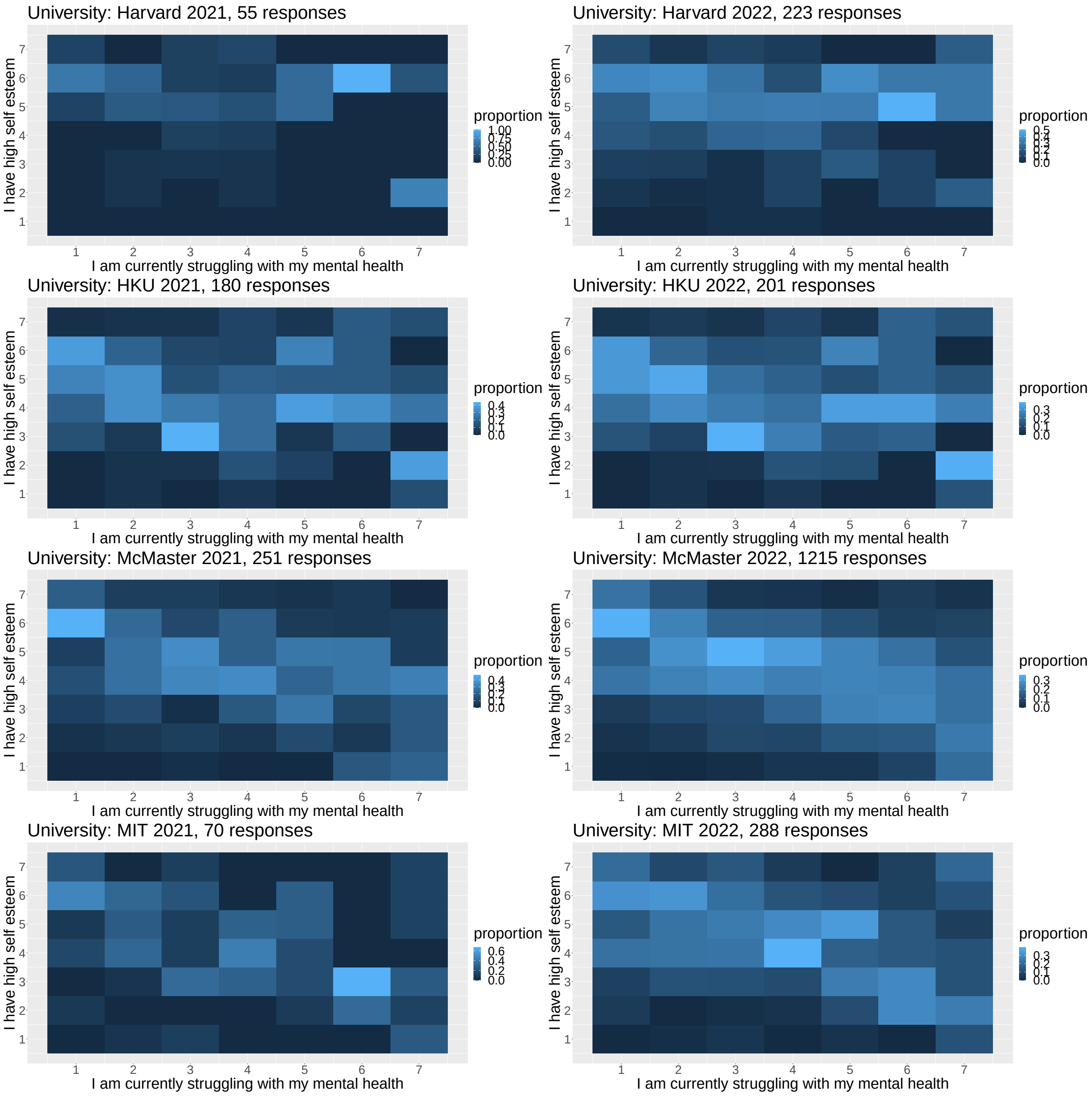


Figure 12: Heatmap of struggling mental health score vs. self esteem

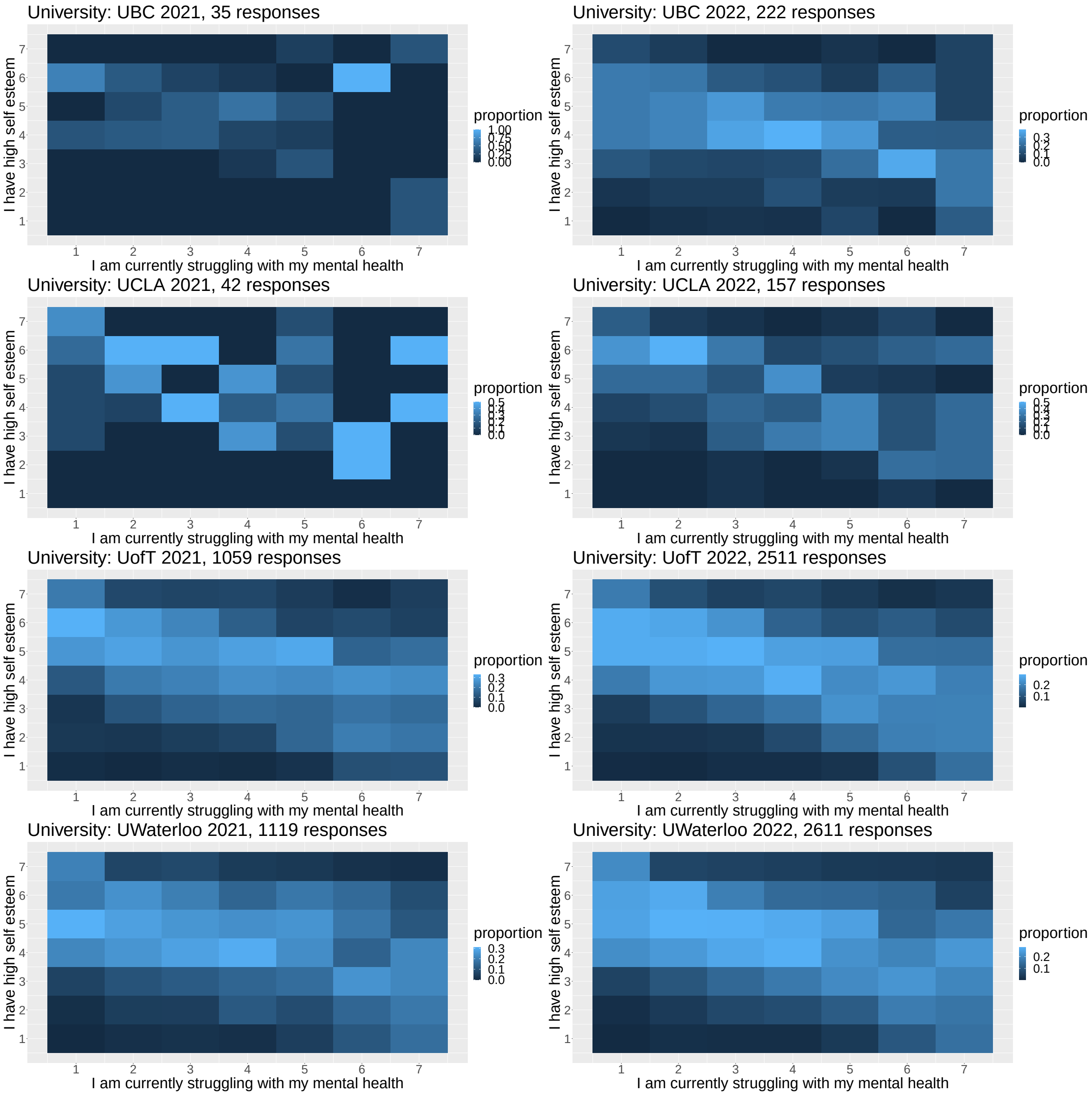
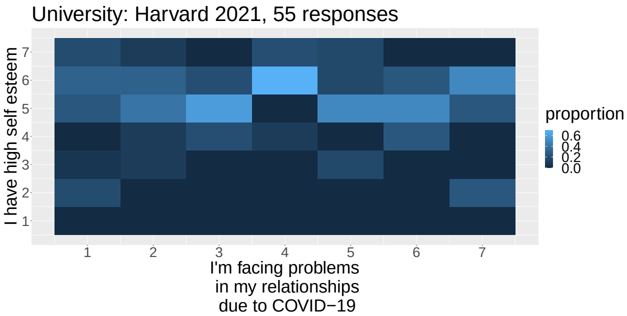
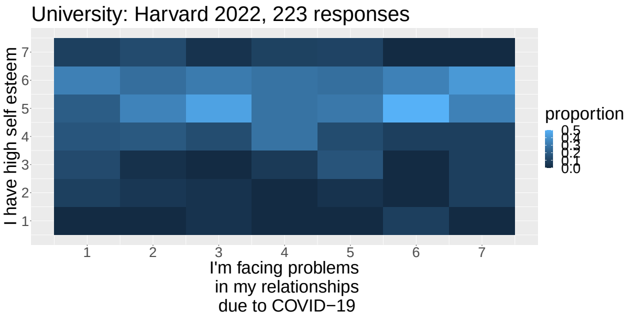
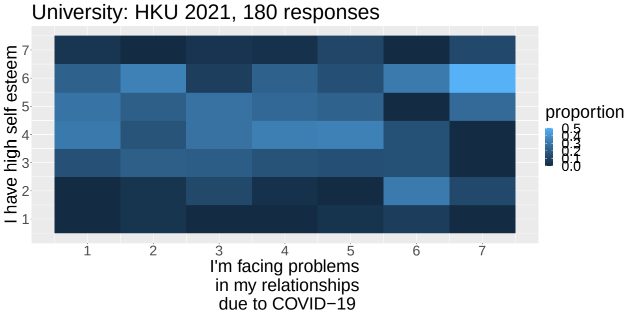
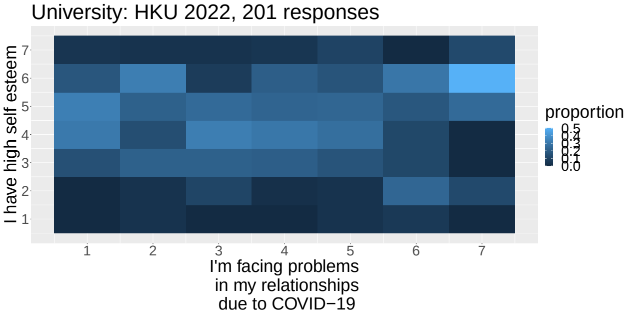


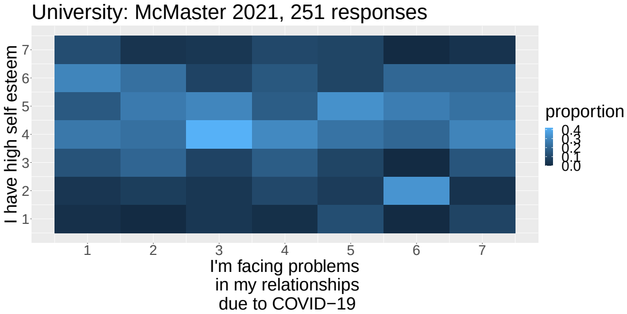
Figure 13: Heatmap of struggling mental health score vs. self esteem

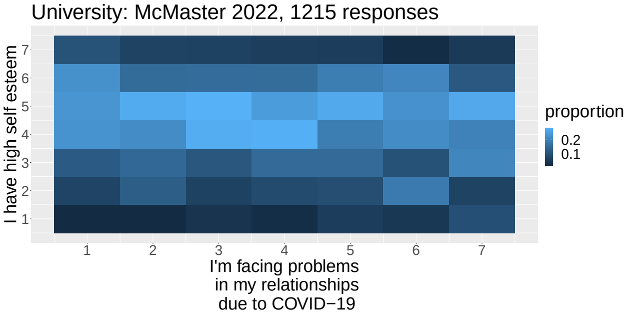

University: Harvard 2021, 55 responses
I have high self esteem
7
6
5
4
3
2
1
1
2
3
4
5
6
7
I'm facing problems
in my relationships
due to COVID−19
proportion
0.6
0.4
0.2
0.0

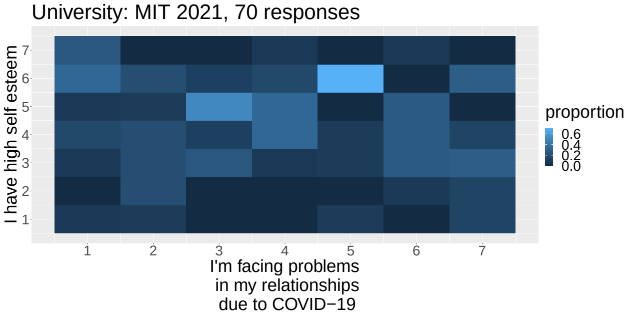

University: Harvard 2022, 223 responses
I have high self esteem
7
6
5
4
3
2
1
1
2
3
4
5
6
7
I'm facing problems
in my relationships
due to COVID−19
proportion
0.5
0.4
0.3
0.2
0.1
0.0

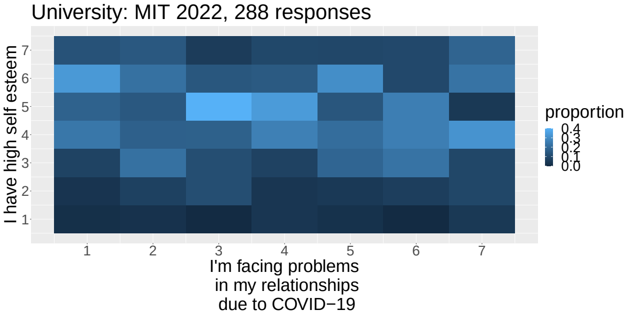

University: HKU 2021, 180 responses
I have high self esteem
7
6
5
4
3
2
1
1
2
3
4
5
6
7
I'm facing problems
in my relationships
due to COVID−19
proportion
0.5
0.4
0.3
0.2
0.1
0.0

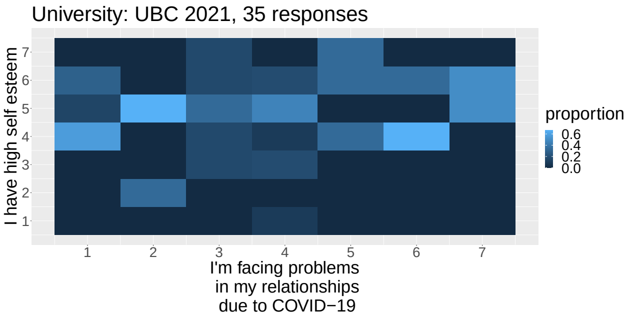

University: HKU 2022, 201 responses
I have high self esteem
7
6
5
4
3
2
1
1
2
3
4
5
6
7
I'm facing problems
in my relationships
due to COVID−19
proportion
0.5
0.4
0.3
0.2
0.1
0.0

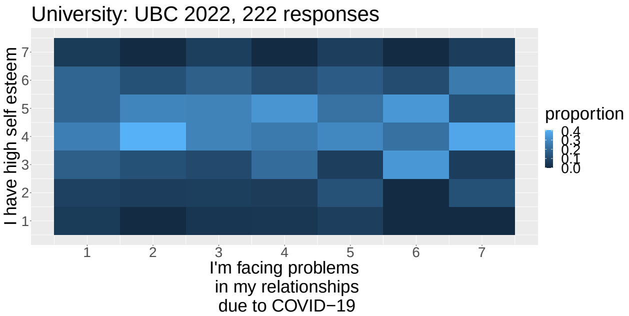

University: McMaster 2021, 251 responses
I have high self esteem
7
6
5
4
3
2
1
1
2
3
4
5
6
7
I'm facing problems
in my relationships
due to COVID−19
proportion
0.4
0.3
0.2
0.1
0.0

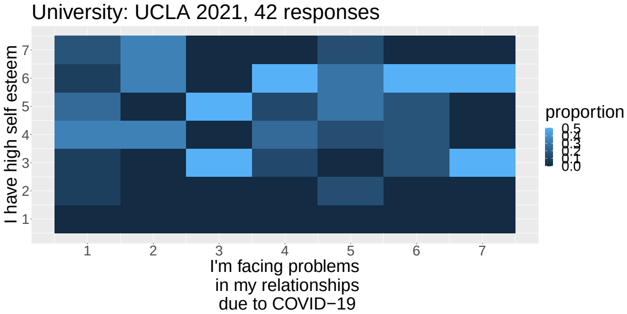

University: McMaster 2022, 1215 responses
I have high self esteem
7
6
5
4
3
2
1
1
2
3
4
5
6
7
I'm facing problems
in my relationships
due to COVID−19
proportion
0.2
0.1

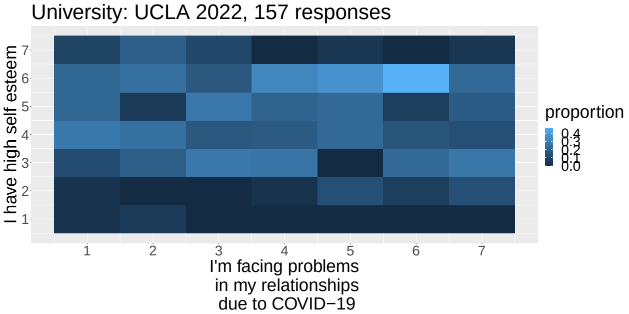

University: MIT 2021, 70 responses
I have high self esteem
7
6
5
4
3
2
1
1
2
3
4
5
6
7
I'm facing problems
in my relationships
due to COVID−19
proportion
0.6
0.4
0.2
0.0

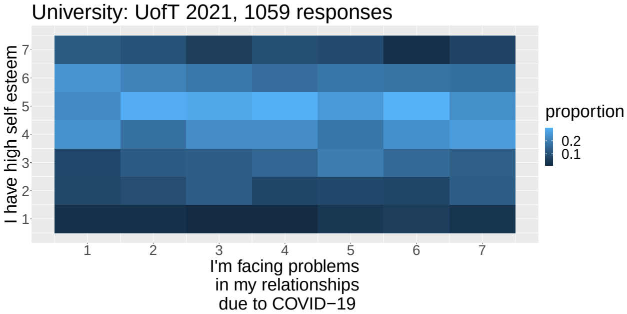

University: MIT 2022, 288 responses
I have high self esteem
7
6
5
4
3
2
1
1
2
3
4
5
6
7
I'm facing problems
in my relationships
due to COVID−19
proportion
0.4
0.3
0.2
0.1
0.0

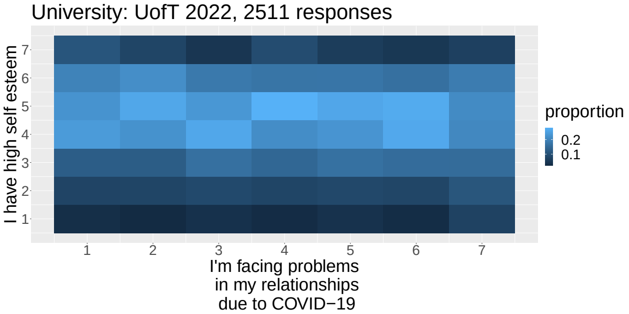


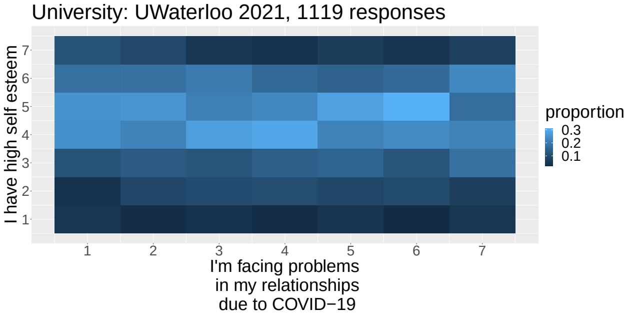


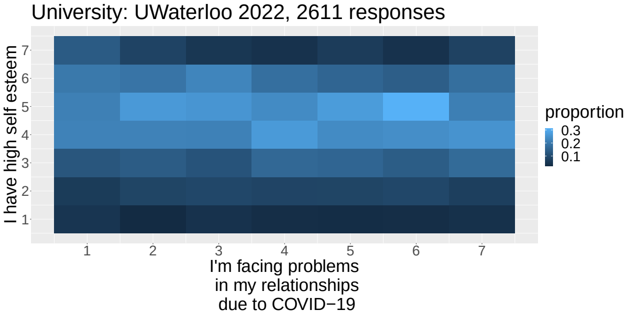












7 Datasheet

Extract of the questions from Gebru et al. (2021)

Motivation

1. *For what purpose was the dataset created? Was there a specific task in mind? Was there a specific gap that needed to be filled? Please provide a description.*
    - The dataset was created for the purpose of matching students together for platonic or romatic relationships. It collects personality data to do so, and matches pairs based off of their responses.
2. *Who created the dataset (for example, which team, research group) and on behalf of which entity (for example, company, institution, organization)?*
    - This dataset was curated by a group of student volunteers organizing the matching service (called the Aphrodite Project).

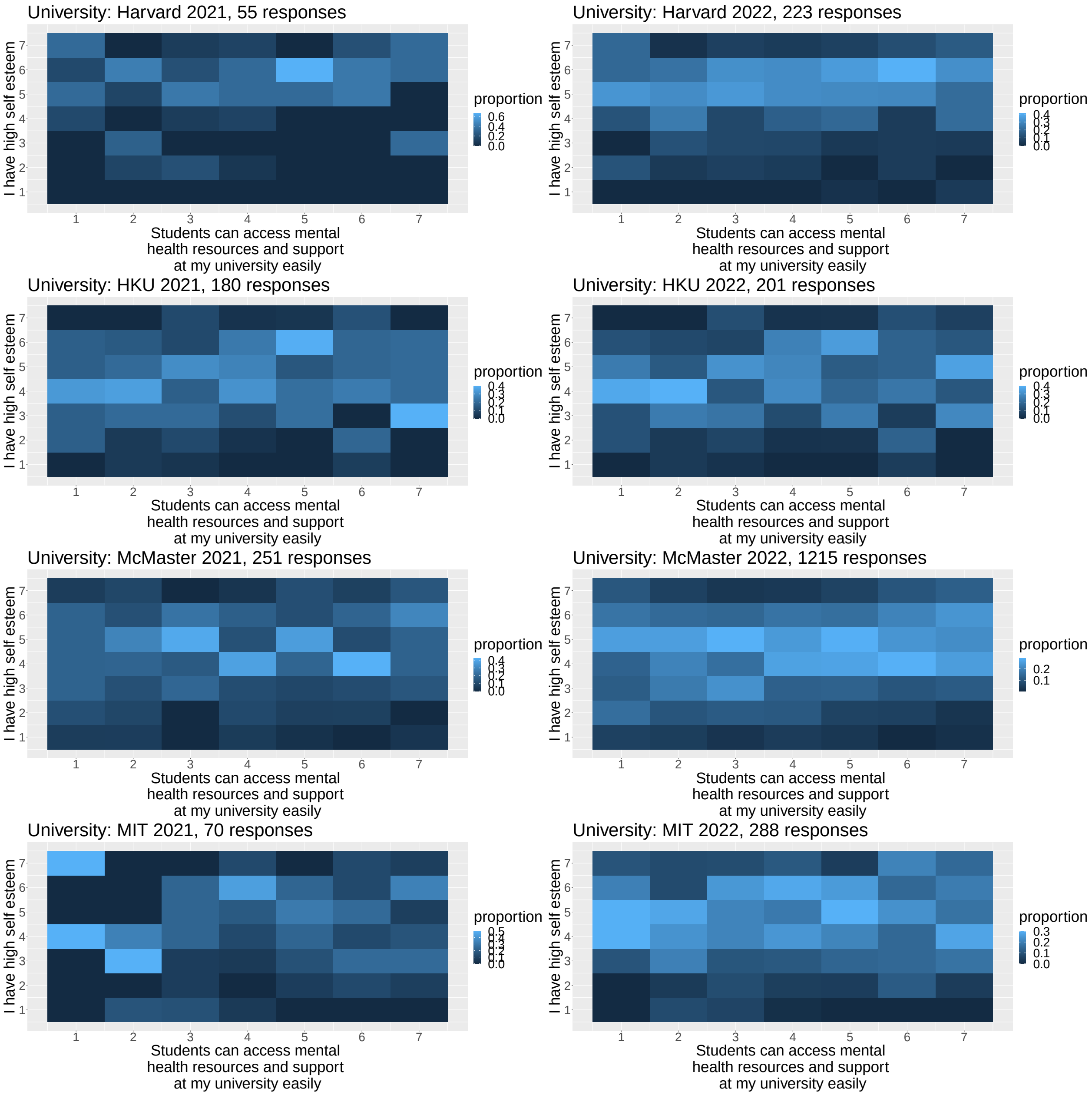


Figure 14: Heatmap of resource accessibility vs. self esteem

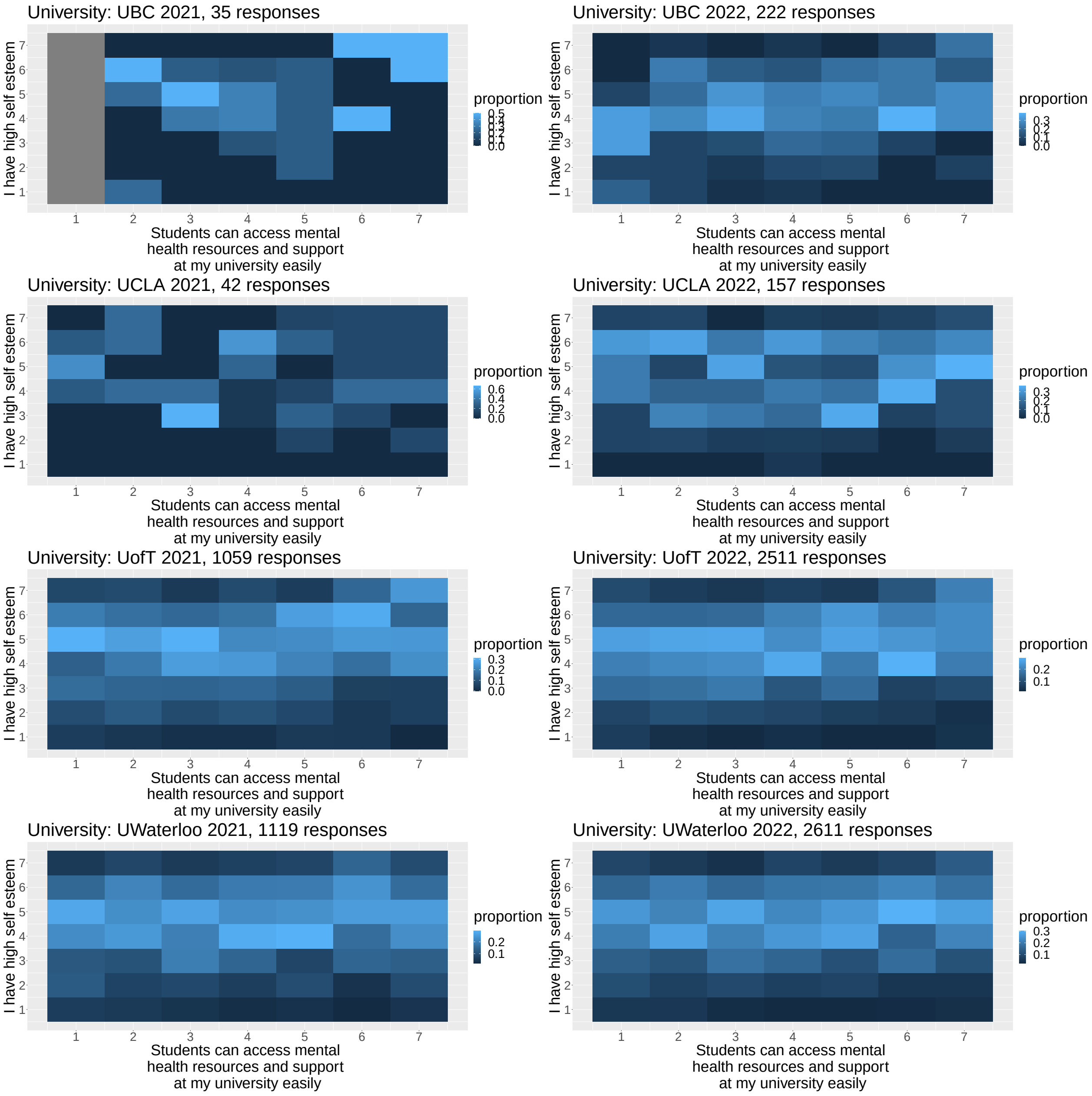


Figure 15: Heatmap of resource accessibility vs. self esteem

3. *Who funded the creation of the dataset? If there is an associated grant, please provide the name of the grantor and the grant name and number.*
   - The Aphrodite Project has received funding in the past from the National University of Singapore, but has received no formal research grants to date.
4. *Any other comments?*
   - N/A

Composition

1. *What do the instances that comprise the dataset represent (for example, documents, photos, people, countries)? Are there multiple types of instances (for example, movies, users, and ratings; people and interactions between them; nodes and edges)? Please provide a description.*
   - Tabular data responses are collected, of personality questions answered on a 1 to 7 Likert scale (from strongly disagree to strongly agree)
2. *How many instances are there in total (of each type, if appropriate)?*
   - There are 76 total personality and optional questions asked to each participant, and counting from 2021 and 2022 there were 13,072 participants.
3. *Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set? If the dataset is a sample, then what is the larger set? Is the sample representative of the larger set (for example, geographic coverage)? If so, please describe how this representativeness was validated/verified. If it is not representative of the larger set, please describe why not (for example, to cover a more diverse range of instances, because instances were withheld or unavailable).*
   - Responses from universities that participated in 2022 but not 2021 are not considered in this analysis, because no analysis around COVID-19 can be done on those schools. There are two schools who fit this bill: Western University and Carnegie Mellon University.
4. *What data does each instance consist of? "Raw" data (for example, unprocessed text or images) or features? In either case, please provide a description.*
   - Each tabular entry is a response to a personality or optional question, where participants can select options using a 1 to 7 Likert scale (from strongly disagree to strongly agree)
5. *Is there a label or target associated with each instance? If so, please provide a description.*
   - Students are assigned a match by the Aphrodite project, however no specific target variable exists
6. *Is any information missing from individual instances? If so, please provide a description, explaining why this information is missing (for example, because it was unavailable). This does not include intentionally removed information, but might include, for example, redacted text.*
   - Optional questions at the end, relating to student mental health, community, and academic life can be missing from individual instances
7. *Are relationships between individual instances made explicit (for example, users' movie ratings, social network links)? If so, please describe how these relationships are made explicit.*
   - Each individual instance is collected independent from others. Similar responses however are matched together and those participants are connected.
8. *Are there recommended data splits (for example, training, development/validation, testing)? If so, please provide a description of these splits, explaining the rationale behind them.*
   - Training and testing splits can be done for modelling with the dataset, however because of small sample sizes in some schools we omitted that process
9. *Are there any errors, sources of noise, or redundancies in the dataset? If so, please provide a description.*
   - There are many possible data biases with this collected data. First, the data is collected as a part of a student matchmaking service, so only students who are comfortable engaging in such a service would have participated. Because of this fact, it is possible that certain demographics of students are entirely excluded from the dataset. For instance, students who are more introverted and prefer not to meet strangers would not participate in the service. Additionally, the mental health questions are asked in the context of matchmaking, which can introduce some unconscious bias.
10. *Is the dataset self-contained, or does it link to or otherwise rely on external resources (for example, websites, tweets, other datasets)? If it links to or relies on external resources, a) are there guarantees that they will exist, and remain constant, over time; b) are there official archival versions of the complete*

dataset (that is, including the external resources as they existed at the time the dataset was created); c) are there any restrictions (for example, licenses, fees) associated with any of the external resources that might apply to a dataset consumer? Please provide descriptions of all external resources and any restrictions associated with them, as well as links or other access points, as appropriate.
    - Because the dataset collects information about student experience and personality, it is more personal than other datasets, and does not extend itself to outside resources
11. *Does the dataset contain data that might be considered confidential (for example, data that is protected by legal privilege or by doctor-patient confidentiality, data that includes the content of individuals' non-public communications)? If so, please provide a description.*
    - No confidential data is collected
12. *Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety? If so, please describe why.*
    - N/A
13. *Does the dataset identify any sub-populations (for example, by age, gender)? If so, please describe how these subpopulations are identified and provide a description of their respective distributions within the dataset.*
    - Yes, demographic questions are asked at the start of the survey. Participants are asked for their age, year of study, race, gender identity, and religious identities.
14. *Is it possible to identify individuals (that is, one or more natural persons), either directly or indirectly (that is, in combination with other data) from the dataset? If so, please describe how.*
    - No, all data is anonymized after the service has been run
15. *Does the dataset contain data that might be considered sensitive in any way (for example, data that reveals race or ethnic origins, sexual orientations, religious beliefs, political opinions or union memberships, or locations; financial or health data; biometric or genetic data; forms of government identification, such as social security numbers; criminal history)? If so, please provide a description.*
    - Yes, questions about religious and political beliefs are asked and responses are stored
16. *Any other comments?*
    - N/A

Collection process

1. *How was the data associated with each instance acquired? Was the data directly observable (for example, raw text, movie ratings), reported by subjects (for example, survey responses), or indirectly inferred/derived from other data (for example, part-of-speech tags, model-based guesses for age or language)? If the data was reported by subjects or indirectly inferred/derived from other data, was the data validated/verified? If so, please describe how.*
    - The data was directly aquired and collected by survey participants
2. *What mechanisms or procedures were used to collect the data (for example, hardware apparatuses or sensors, manual human curation, software programs, software APIs)? How were these mechanisms or procedures validated?*
    - Participants filled out the survey on an online website, which stored the data in a database on the cloud
3. *If the dataset is a sample from a larger set, what was the sampling strategy (for example, deterministic, probabilistic with specific sampling probabilities)?*
    - N/A
4. *Who was involved in the data collection process (for example, students, crowdworkers, contractors) and how were they compensated (for example, how much were crowdworkers paid)?*
    - Unpaid student volunteers were involved in curating the dataset
5. *Over what timeframe was the data collected? Does this timeframe match the creation timeframe of the data associated with the instances (for example, recent crawl of old news articles)? If not, please describe the timeframe in which the data associated with the instances was created.*
    - The dataset was curated over 2 years, from January 2021 to February 2022.
6. *Were any ethical review processes conducted (for example, by an institutional review board)? If so, please provide a description of these review processes, including the outcomes, as well as a link or other access point to any supporting documentation.*

- N/A

7. *Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (for example, websites)?*
   - Yes, I collected the data
8. *Were the individuals in question notified about the data collection? If so, please describe (or show with screenshots or other information) how notice was provided, and provide a link or other access point to, or otherwise reproduce, the exact language of the notification itself.*
   - Yes, individuals were required to sign up, and agreed to a term of service to having data analyzed in aggregate (but to have all responses anonymized so no identification could occur)
9. *Did the individuals in question consent to the collection and use of their data? If so, please describe (or show with screenshots or other information) how consent was requested and provided, and provide a link or other access point to, or otherwise reproduce, the exact language to which the individuals consented.*
   - Yes, a terms and service was agreed to at the start of the survey
10. *If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or for certain uses? If so, please provide a description, as well as a link or other access point to the mechanism (if appropriate).*
    - Yes, participants may request to delete all of their data on the Aphrodite project website at any time
11. *Has an analysis of the potential impact of the dataset and its use on data subjects (for example, a data protection impact analysis) been conducted? If so, please provide a description of this analysis, including the outcomes, as well as a link or other access point to any supporting documentation.*
    - N/A
12. *Any other comments?*
    - N/A

Preprocessing/cleaning/labeling

1. *Was any preprocessing/cleaning/labeling of the data done (for example, discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)? If so, please provide a description. If not, you may skip the remaining questions in this section.*
   - No, only Likert responses are collected, so no preprocessing is necessary
2. *Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data (for example, to support unanticipated future uses)? If so, please provide a link or other access point to the "raw" data.*
   - No
3. *Is the software that was used to preprocess/clean/label the data available? If so, please provide a link or other access point.*
   - No, it is closed-source
4. *Any other comments?*
   - N/A

Uses

1. *Has the dataset been used for any tasks already? If so, please provide a description.*
   - Yes, it has been used to match students based off of their personality responses using the Stable Roommates matching algorithm
2. *Is there a repository that links to any or all papers or systems that use the dataset? If so, please provide a link or other access point.*
   - No
3. *What (other) tasks could the dataset be used for?*
   - Analyses related to student mental health and student communities could be conducted using this dataset, as demographic information as well as personality information is collected
4. *Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses? For example, is there anything that a dataset consumer might need to know to avoid uses that could result in unfair treatment of individuals or groups (for example, stereotyping, quality of service issues) or other risks or harms (for example, legal risks, financial harms)? If so, please provide a description. Is there anything a dataset consumer could do to*

mitigate these risks or harms?
  - Algorithm bias certainly exists, and it is the responsibility of the service to be aware of this. However, because the service is opt-in, and participants are able to disconnect with their match and not reveal their contact information, these consequences are minimized
5. *Are there tasks for which the dataset should not be used? If so, please provide a description.*
    - The dataset should not be used to identify vulnerable student groups for exploitation. It could be possible to do so, since demographic and personality data is collected.
6. *Any other comments?*
    - N/A

Distribution

1. *Will the dataset be distributed to third parties outside of the entity (for example, company, institution, organization) on behalf of which the dataset was created? If so, please provide a description.*
    - No future plans for this to happen
2. *How will the dataset be distributed (for example, tarball on website, API, GitHub)? Does the dataset have a digital object identifier (DOI)?*
    - N/A
3. *When will the dataset be distributed?*
    - N/A
4. *Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)? If so, please describe this license and/ or ToU, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms or ToU, as well as any fees associated with these restrictions.*
    - N/A
5. *Have any third parties imposed IP-based or other restrictions on the data associated with the instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms, as well as any fees associated with these restrictions.*
    - N/A
6. *Do any export controls or other regulatory restrictions apply to the dataset or to individual instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any supporting documentation.*
    - N/A
7. *Any other comments?*
    - N/A

Maintenance

1. *Who will be supporting/hosting/maintaining the dataset?*
    - The Aphrodite project takes responsibility for maintaining the dataset
2. *How can the owner/curator/manager of the dataset be contacted (for example, email address)?*
    - Participants can contact the team through the Aphrodite project website
3. *Is there an erratum? If so, please provide a link or other access point.*
    - N/A
4. *Will the dataset be updated (for example, to correct labeling errors, add new instances, delete instances)? If so, please describe how often, by whom, and how updates will be communicated to dataset consumers (for example, mailing list, GitHub)?*
    - Yes, the dataset gets updated every run of the Aphrodite project, which happens annually
5. *If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (for example, were the individuals in question told that their data would be retained for a fixed period of time and then deleted)? If so, please describe these limits and explain how they will be enforced.*
    - Individuals are told that their data will be anonymized immediately after the service concludes, however anonymized data may be kept
6. *Will older versions of the dataset continue to be supported/hosted/maintained? If so, please describe how. If not, please describe how its obsolescence will be communicated to dataset consumers.*

- No plans as of yet

7. *If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so? If so, please provide a description. Will these contributions be validated/verified? If so, please describe how. If not, why not? Is there a process for communicating/distributing these contributions to dataset consumers? If so, please provide a description.*
   - No plans as of yet
8. *Any other comments?*
   - N/A

8 References

n.d. https://sites.psu.edu/ccmh/files/2017/10/2015_CCMH_Report_1-18-2015-yq3vik.pdf.

Arel-Bundock, Vincent. 2022. *Modelsummary: Summary Tables and Plots for Statistical Models and Data: Beautiful, Customizable, and Publication-Ready*. https://CRAN.R-project.org/package=modelsummary.

Blanco, Carlos, Mayumi Okuda, Crystal Wright, Deborah S. Hasin, Bridget F. Grant, Shang-Min Liu, and Mark Olfson. 2008. "Mental Health of College Students and Their Non-College-Attending Peers: Results from the National Epidemiologic Study on Alcohol and Related Conditions." *Archives of General Psychiatry* 65 (12): 1429–37. https://doi.org/10.1001/archpsyc.65.12.1429.

"Covid-19: Student Survey." 2021. *Active Minds*. https://www.activeminds.org/studentsurvey/.

Eisenberg, Daniel, Ezra Golberstein, and Justin B Hunt. 2009. "Mental Health and Academic Success in College." *The B.E. Journal of Economic Analysis & Policy* 9 (1). https://doi.org/doi:10.2202/1935-1682.2191.

Firke, Sam. 2021. *Janitor: Simple Tools for Examining and Cleaning Dirty Data*. https://CRAN.R-project.org/package=janitor.

Gebru, Timnit, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé Iii, and Kate Crawford. 2021. "Datasheets for Datasets." *Communications of the ACM* 64 (12): 86–92.

Gilburt, Helen, Diana Rose, and Mike Slade. 2008. "The Importance of Relationships in Mental Health Care: A Qualitative Study of Service Users' Experiences of Psychiatric Hospital Admission in the UK." *BMC Health Services Research* 8 (1): 92. https://doi.org/10.1186/1472-6963-8-92.

Liaw, Andy, and Matthew Wiener. 2002. "Classification and Regression by randomForest." *R News* 2 (3): 18–22. https://CRAN.R-project.org/doc/Rnews/.

Markoulakis, Roula, and Bonnie Kirsh. 2013. "Difficulties for University Students with Mental Health Problems: A Critical Interpretive Synthesis." *The Review of Higher Education*. Johns Hopkins University Press. https://muse.jhu.edu/article/522572/summary.

Matias, Thiago, Fabio H Dominski, and David F Marks. 2020. "Human Needs in COVID-19 Isolation." *Journal of Health Psychology* 25 (7): 871–82. https://doi.org/10.1177/1359105320925149.

R Core Team. 2021. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Sonnak, Carina, and Tony Towell. 2001. "The Impostor Phenomenon in British University Students: Relationships Between Self-Esteem, Mental Health, Parental Rearing Style and Socioeconomic Status." *Personality and Individual Differences* 31 (6): 863–74. https://doi.org/https://doi.org/10.1016/S0191-8869(00)00184-7.

Wickham, Hadley. 2016. *Ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. https://ggplot2.tidyverse.org.

———. 2021a. *Forcats: Tools for Working with Categorical Variables (Factors)*. https://CRAN.R-project.org/package=forcats.

———. 2021b. *Tidyr: Tidy Messy Data*. https://CRAN.R-project.org/package=tidyr.

Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, et al. 2019. "Welcome to the tidyverse." *Journal of Open Source Software* 4 (43): 1686. https://doi.org/10.21105/joss.01686.

Wickham, Hadley, Romain François, Lionel Henry, and Kirill Müller. 2021. *Dplyr: A Grammar of Data Manipulation*. https://CRAN.R-project.org/package=dplyr.