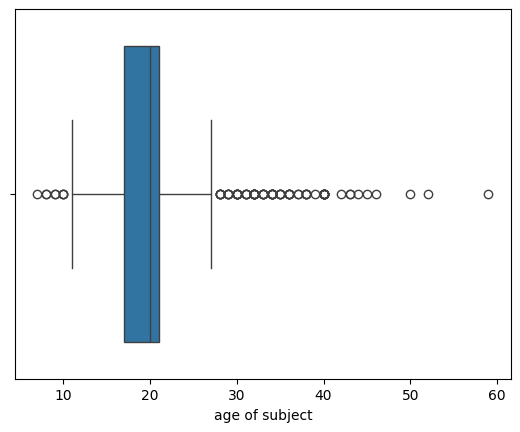
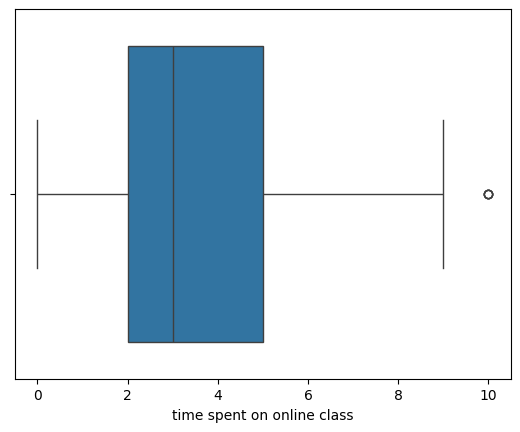
**Group members**: Grace Brown (fth7te) (GitHub: graceabrown12), Nicholas Hughes (zdw3mp) (GitHub: nicholas99212), Zoya Masood (rpk4wp) (GitHub: zoyamasood), Emma Mills (Github: EmmaMills1002; zxy9ss), Bora Ya Diul (ngx3fy, git user - borayd3), Lauren Wisniewski (zxf3df, GitHub: laurenwisniewski), Valyn Grebe (eyv7jz) (GitHub: eyv7jz)

**Main question**:

Our main question asks how mental health was impacted by the Covid-19 pandemic, particularly for students who suddenly moved to online classes. Because our dataset surveys students located in India, our results will be specific to their context. To do so, the dependent variables that we will investigate will serve as indicators of a student’s mental health, including time spent on wellness (sleep, fitness), health changes (weight, meals, health issues), and how students managed stress. The independent variable we will look at is age, which we grouped into age groups for more clear analysis. The mean and median age of our participants are both 20, with 50% of participants between 17 and 21.



We will also look at the number of time spent on online classes as a predictor, where students indicated the number of hours they spent taking online classes, with answers ranging from zero to ten hours.



**What is an observation in your study?**

An observation in our study represents an individual student, who self reports the impacts of the Covid-19 pandemic on their academics and overall wellbeing. Using proxy variables to represent mental health will help us gauge how students were impacted without directly asking them to classify their own mental health. For example, we expect that low sleep hours will indicate poor mental health, whereas higher sleep hours will indicate good mental health.

**What models or algorithms do you plan to use in your analysis? How?**

For our data, we plan to use Multiple Linear Regression. As mentioned, we will use the proxy variables for evaluating mental health in students, such as “time spent on sleep”, “time spent on fitness”, “changes in weight”, “number of meals per day”, “health issues during lockdown”, and the ways students mediated their stress as our response variables, respectively, and use explanatory variables such as “age”, and “time spent on online class” as predictors. Within our MLR, we will also employ a train test split to simulate running our regression on novel data through the sklearn library in Python. With this train test split, we will utilize an 80 train and 20 test split on our data.

Within our regression equation, we will also pay special attention to any variables that should be transformed based on our EDA and include those transformed variables within our separate regressions for the predictors..

**Are you doing supervised or unsupervised learning? Classification or regression?**

Because we plan on using Multiple Linear Regression, we will be performing supervised learning. Our data will already have labels, and we will be discovering the relationship between the different classes of our independent variables and our dependent variables. We will also be performing a regression rather than a classification, and quantifying the effect of our independent variables numerically.

**How will you know if your approach “works”? What does success mean?**

We have determined that our approach will “work” if our model produces performance metrics, such as R squared and RMSE, which we can later evaluate for success. We have determined this as our base standard for knowing our approach “works” as our model is able to predict whether age affects mental health indicators, such as wellness, health changes, and stress levels. However, the accuracy of these predictions is what we have determined as predicting success. We would like a high R squared value which would indicate that a significant proportion of the variance in mental health indicators is explained by either age or time in online classes (our predictors). Additionally, a lower RMSE would indicate the model’s predictions are similar to the actual observations (with minimal errors). To us, success means that our model achieves a balance between a high R squared and low RMSE. We expect the model to be robust, showing similar performance on training and test sets. Additionally, success would mean we are able to interpret the model’s coefficients indicating meaningful and expected relationships. For example, we would expect more hours of sleep, on average, to indicate some mental health concerns as this is consistent with current research.

**What are weaknesses that you anticipate being an issue? How will you deal with them if they come up? If your approach fails, what might you learn from this unfortunate outcome?**

We anticipate that there might be some overlap in how our independent variables interact to create an effect on the dependent variable. If this issue comes up, we will deal with this by creating interaction features/variables through transformations. For example, we might end up running our regression of sleep time on both age and online class time after combining these two variables into an interaction variable. If this approach fails, we will learn that there may be a more complicated relationship between our chosen variables and other variables in the dataset. To deal with this, we may need to use a technique like PCA to figure out which variables truly explain most of the variation in our data.

**Feature Engineering: How will you prepare the data specifically for your analysis? For example, are there many variables that should be one-hot encoded? Do you have many correlated numeric variables, for which PCA might be a useful tool?**

To prepare the data specifically for analysis, the categorical variables of age group will be one-hot encoded to allow our model to interpret these values without misleading a numerical relationship. Age can be grouped in intervals of high school and college, with high school being 14-18 and college being 18-22. This will provide straightforward analysis that is differential dependent on school placement. Interaction may exist between time in online classes and age group. If there is collinearity observed, we can create interaction variables here to make sure the effects from each are combined. For time spent online, sleep time, and wellness time, values can be normalized to interpret coefficients consistently. This will also prevent any variable from dominating the overall regression due to its scale.

PCA may be useful for highly correlated variables, such as wellness, sleep, and health issues. PCA can be applied if needed to reduce overall dimensionality and create principal components that will be capable of capturing variance and avoiding multicollinearity in the regression. Lastly, missing values can be checked for based on frequency and overall affected variables. This will help us ensure that our model is not biased from incomplete data.

**Results: How will you communicate or present your results? This might be a table of regression coefficients, a confusion matrix, or comparisons of metrics like $R^2$ and RMSE or accuracy and sensitivity/specificity. This is how you illustrate why your plan succeeded or explain why it failed.**

To effectively communicate the results, we will use visualization and tables to present our findings from our multiple linear regression to support our conclusions. For visualization, we will utilize residual plots, which will be helpful in assessing the assumptions of linear regression by checking for homoscedasticity and the normality of our residuals. Residual plots are good for determining model sensitivity by showing highly leveraged data points. For the tables, we will have one for the metrics R squared and one for RMSE, which will be calculated based on the regression coefficients derived from the model. A side-by-side comparison of these metrics for both the training and test datasets will help illustrate the robustness of our model.

Data: <https://www.kaggle.com/datasets/kunal28chaturvedi/covid19-and-its-impact-on-students>

GitHub project repo: <https://github.com/laurenwisniewski/DS-3001-Project>