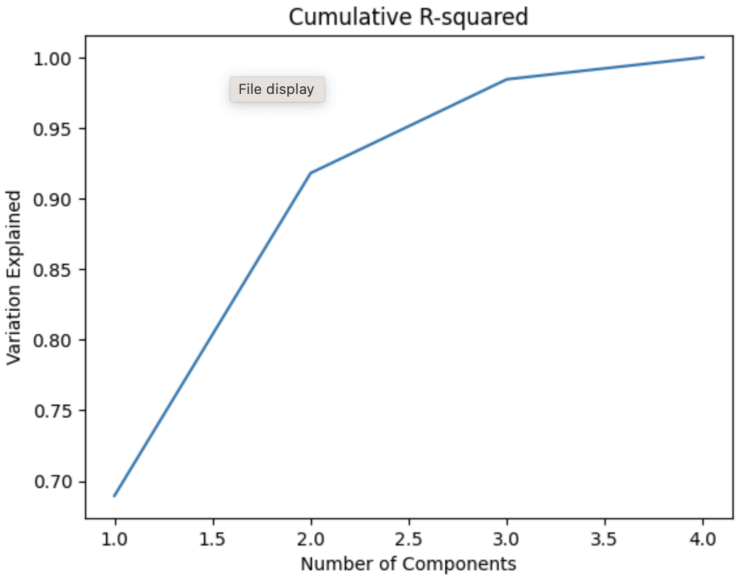
Group members: Grace Brown (fth7te) (GitHub: graceabrown12), Nicholas Hughes (zdw3mp) (GitHub: nicholas99212), Zoya Masood (rpk4wp) (GitHub: zoyamasood), Emma Mills (Github: EmmaMills1002; zxy9ss), Bora Ya Diul (ngx3fy, git user - borayd3), Lauren Wisniewski (zxf3df, GitHub: laurenwisniewski), Valyn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

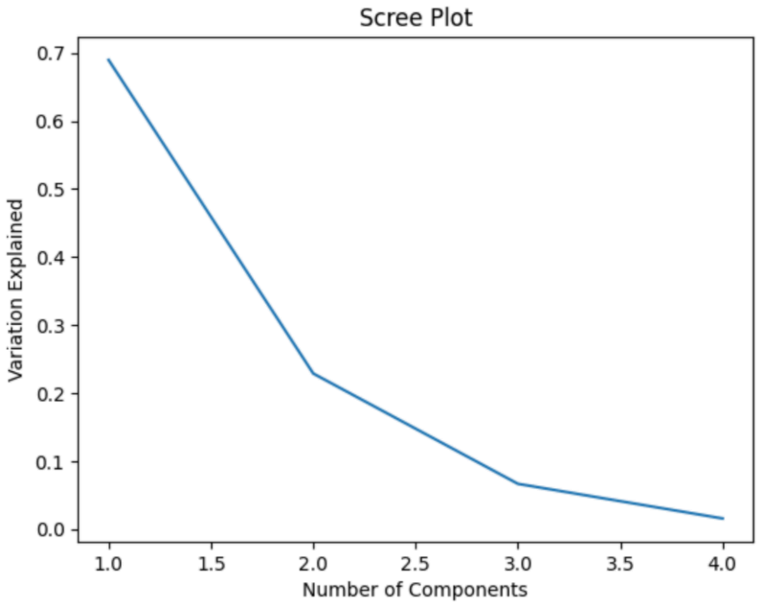
Data: https://www.kaggle.com/datasets/kunal28chaturvedi/covid19-and-its-impact-on-students
GitHub project repo: https://github.com/laurenwisniewski/DS-3001-Project

Our main question was to see whether we could use lifestyle variables to predict an important indicator of health, which could help identify which students are more likely to need more support at school. The lifestyle variables we considered were the age of students, time spent on social media, number of meals per day, and the combined time spent on online class as well as independent study (the interaction variable classAndStudy). The health variable we used was time spent on sleep. We believe that a predictive model such as ours (which uses lifestyle variables to predict an important measure of health in students) could be useful to university health clinics who may want to identify individuals at a potentially higher risk for struggling with their physical or mental health. As will be noted in our criticisms and concerns section, however, our model is based on correlations, not causations and should be used with caution.

To address our prediction question, we first did a PCA decomposition of the lifestyle variables to find that the first three principal components explained almost all of the variation in our dataset. Then, we ran linear regression using these three components to find small but significant correlations between the lifestyle variables and our health variable. Our R-squared value was 0.09 and the correlation coefficients for our lifestyle variables were -0.02091545 (for age), -0.07014782 (for the classAndStudy variable), 0.00909555 (for time spent on social media), and 0.12185532 (for number of meals per day). These correlation coefficients can be used to predict hours of sleep (a representation of overall student wellbeing in our study) from our lifestyle variables.

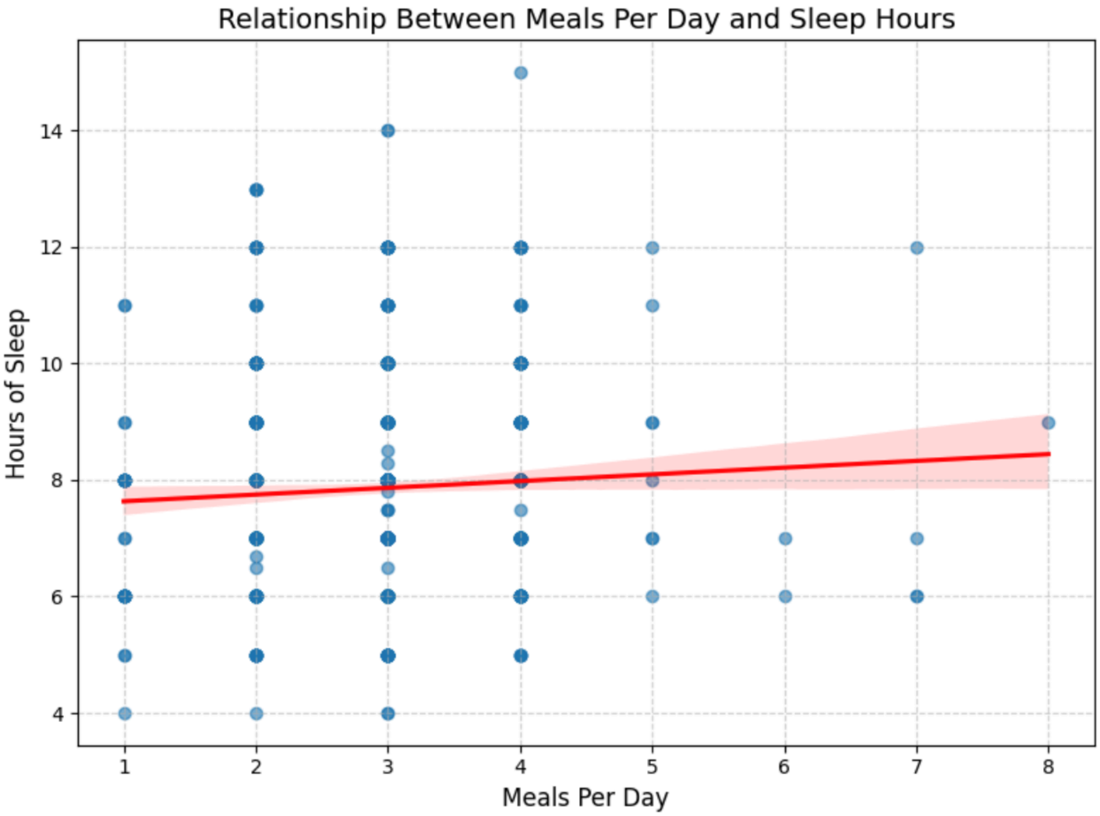


The above graph illustrates the cumulative R-squared as a function of the number of components, showing the percentage of variation explained as additional components are added. There is a significant increase in cumulative R-squared from one to two components, with explained variation rising from approximately 70% to over 90%. Adding a third component increases the explained variation to over 95%, while the fourth component brings it close to 100%. Using only one component in our PCA would likely result in underfitting, as it fails to capture sufficient variance in the data. Conversely, using all four components may lead to overfitting, where the model becomes overly complex and less generalizable. Selecting two components provides an adequate balance, capturing around 90% of the variance. However, we wanted a higher threshold of explained variance closer to 95%, so incorporating a third component was preferable. This graph was helpful determining the optimal number of components to use in the principal component analysis, ensuring a model that balances simplicity with explanatory power.



The Scree Plot above demonstrates the eigenvalues of the PCA components within our analysis and looks at how much of the variance can be explained by a certain number of components. When looking at this graph, there is a large drop from one to two components before the decline is a bit more gradual from two components to four components. At one component, the corresponding eigenvalue of 0.689 before steeply dropping off to an eigenvalue of 0.229 when a second component is added. With a third component, the amount of variance explained decreases to a value of .066 and the addition of a fourth component drops this amount to 0.0156. The 'elbow' point of this graph, or where the variance explained by the number of components begins to level off, appears to occur when there are three components included. This is significant as it indicates that including one to two components within our PCA would likely result in an optimal output for our analysis. There is the initial drop in the amount of variance explained from one to two components, however, this drop off continues until three components where the eigenvalues begin to level off. Therefore, utilizing only one or two components will not optimally explain the variance whereas using three components will. Further, using a number of components past three will result in overfitting the data, as it is past the optimal level.

Because our dataset is a cross-sectional survey of students taken at a single point in time, we do not recommend drawing causal relationships from our linear regression model. For example, we found that the number of meals per day has a positive relationship with our proxy variable for health, time spent on sleep. This does suggest that students who eat a lesser number of meals per day are more likely to have worse mental health statuses, however it does not imply that increasing the number of these students' meals would automatically lead to better health outcomes.



The scatterplot examining the relationship between meals per day and hours of sleep reveals a slight positive trend, where individuals consuming more meals tend to sleep marginally longer. However, the weak slope of the regression line and the wide confidence interval suggest a low correlation, indicating that meals per day alone do not strongly predict sleep patterns. Most observations cluster around 1-3 meals per day, with sleep hours ranging widely within this group. This finding highlights the limited role of meal frequency in explaining sleep duration and suggests that additional variables, such as stress levels, time spent online, or physical activity, may also account for variations in sleep behavior. This visualization underscores the importance of considering multiple factors when modeling behaviors influenced by complex health and lifestyle interactions.