Group members: Gracie 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)

Our dataset contains information about COVID-19 and its impact on students attending educational institutions in the Delhi National Capital Region (NCR) in India. The data were gathered through a cross-sectional survey which was completed by 1,182 students of different age groups in the NCR area in 2020. The original dataset contained rows for each respondent and their answers to the survey. The columns (and their data type) in the original dataset were as follows: region of residence (str), age of subject (int), time spent on online class (float), rating of online class experience (str), medium for online class (str), time spent on self study (float), time spent on fitness (float), time spent on sleep (float), time spent on social media (float), preferred social media platform (str), time spent on TV (str), number of meals per day (int), change in your weight (str), health issue during lockdown (str), stress busters (str), time utilized (str), do you find yourself more connected with your family, close friends, and relatives ? (str), what you miss the most (str).

For our cleaning process, we dropped 92 responses that had at least one missing value in one or more of the “rating of online class experience,” “medium for online class”, or “preferred social media platform” columns. We also changed the column names to lowercase to account for differences in capitalization in the original dataset. We then condensed the “stress busters” and “what you miss the most” responses to fit into more generalized categories (18 and 13 categories, respectively). This was useful so that we could see counts of each category to get a more generalized view of what students missed the most and what coping mechanisms they used a lot during the pandemic. The overall cleaning process presented few challenges, as the survey contained high-quality data that only required slight modifications.

This dataset is particularly useful as it sheds light on the multifaceted impact of COVID-19 on students' experiences with virtual learning. Studying this dataset provides insight into how the COVID-19 lockdown affected students’ routines, mental health, and overall education in the

Delhi National Capital Region. By analyzing data surrounding virtual school, stress levels, studying, social media, and physical activity of students, we can identify the specific challenges students faced during lockdown. This information will enable researchers to better understand the factors that influence students' performance, lifestyle, and health trends, ultimately improving educational strategies going forward. Policymakers can use this data to identify key areas for improvement in online education, such as integrating more engaging learning environments and catering toward students' preferred learning mediums. Understanding the stressors and weaknesses in academic performance from virtual learning environments is crucial for enhancing virtual learning frameworks and ensuring they meet the diverse needs of students in future educational scenarios.

Based on our EDA, we found that many causal relationships were based on numerous COVID-19 related factors occurring simultaneously. This made it challenging to isolate a single factor and predict how one pandemic-related variable could affect students' overall well-being or school performance. For example, it would be difficult to measure whether isolation due to the pandemic impacted students' academic performance, as changes in performance could also be attributed to distance learning. As a result, we shifted our focus to investigating how age or educational level could impact mental health.

We immediately noticed the majority of the data are from students aged 13-23 years old. In India, depending on the program, university can last between 3-5 years. Considering this, we grouped students by level of education: university students (ages 18-23) and high school (secondary) students (ages 13-17). Additionally, since we do not have a variable that explicitly measures mental health, we used proxy variables such as "Do you find yourself more connected to family and friends?", "time spent on sleep," and "number of meals per day," etc. We chose these proxy variables because they reflect the physical effects of mental health.

In addition to the challenge of identifying predictive variables, we anticipate difficulties in applying our findings to other geographical regions and school environments outside of the NCR in India. Since, the survey performed an in-depth questionnaire of the virtual learning experience for NCR students, our predictive model will be able to accurately predict how educational level

affects mental health for NCR students. However, there will be significant challenges in applying our insights to different learning environments and cultural contexts.

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