Post-pandemic there has been a significant transition in the workforce from primarily on-site work to remote and hybrid work. One detrimental effect of social isolation, brought on by the pandemic, is poor mental health outcomes. Given that large institutions such as businesses, healthcare and education systems have continued aspects of remote work, it is important to explore the effects on employee mental health. Stemmed from our curiosity, our group chose to use the ["Impact of Remote Work on Mental Health"](https://www.kaggle.com/datasets/waqi786/remote-work-and-mental-health/data) dataset from Kaggle. This dataset explores factors that influence employees' work-life balance, mental health, and other work-related metrics.

Our dataset consisted of 20 columns and 5,000 rows. The data cleaning process began with assessing duplicate entries, which confirmed that no duplicate rows existed in the dataset. We also dropped irrelevant columns, including the Employee\_ID column, which provided no analytical value. We continued by searching for missing values, which yielded over a thousand missing values in both the Physical\_Activity and Mental\_Health\_Condition columns. Our group surmised that a blank answer in these instances was equivalent to “None,” which we imputed for each column. We finally created pair plots to identify potential outliers, but due to the nature of our data, identified none. Many of the numeric columns are on a scale of 1-5, except for things like age or years of experience. As a result, our data was very uniform in its distribution and offered no outliers. Overall, these steps ensured a robust and clean dataset for subsequent analysis and modeling.

While creating the random forest regressor, I encountered a few problems that prevented my code from executing correctly. The first significant issue was my selection of variables. Selecting columns such as Physical\_Activity, Sleep\_Quality, Access\_to\_Mental\_Health\_Resources, and Mental\_Health\_Condition, I had to convert these categorical values to numerical values. However, even after this conversion I ran into a ValueError. Specifically for my predictor variable, Mental\_Health\_Condition, I was receiving the ValueError that NaN was detected in the column. To resolve this problem, I handled the potentially missing values or unsuccessfully conveyed values by assigning these values with 0. After implementing these changes, my code successfully calculated the RSME.

Creating logistic regression for predicting work life balance rating, some problems I encountered was having issues with the target variable. What I decided to do was make the target variable, Work\_Life\_Balance\_Rating, be converted into binary for an example 1 would be "good" ratings of 4 or 5, and 0 for "Poor" ratings of 1,2, or 3). I've also changed the numbers and noticed this setting would have the highest accuracy. The data was split into 80% training and 20% testing. For my visualization I created a confusion matrix and bar chart which helped me show the difference between "Poor" and "Good" work life balance ratings. However by changing the numbers around during the conversion I found the highest accuracy result I could get was 0.54. We've also created a multiple linear regression however we found that the r-squared result was zero which meant the model explained very little of the variability in the target variable.

**2)** [Google Colab Notebook](https://colab.research.google.com/drive/1RYnFuIy_vK_Fo0s1TmbuskmU7cDrSH9k?usp=sharing)

**3)** [MIS515 Group 11 Project](https://drive.google.com/drive/folders/1Up-oNQ6nuxow7iWyTLq8rRlgrAiBW1Gc?usp=sharing)

**4)** <https://drive.google.com/file/d/1f4f-9KYAYoFeZbSHHKFcxlNs52vHfr3F/view?usp=sharing>