**Student Health Data Examination**

In this project, we examined a dataset which recorded student health information. This dataset was interesting to all of us because we have all recently become students once again as part of the data science bootcamp. Our goal with this data was to examine and analyze what health-related factors correlated with others so that we could provide observations and suggestions for students to be successful.

We retrieved this dataset from Kaggle as a CSV file. It was a relatively clean data set to begin with, and we did not need to drop any rows or columns. This data set did not provide insight as to the data collection methods that were used. Our initial data cleaning steps involved importing the CSV into a Jupyter Notebook and creating a basic Pandas data frame which we could all use as a starting point to manipulate this data. In addition to this data frame which we all used as a starting point for our individual code, we also did some data cleaning specific to the individual questions we worked one.

The first line of investigation was to determine what effect stress had on overall health risk among students. We first created a new data frame that contained stress related variables such as blood pressure, heart rate and amount of time spent studying. This data set also included self-reported and biosensor measure stress levels, which were included in this stress related data frame. As an additional step, we had to convert qualitative data such as the a “Low, Medium, or High” health risk into something quantitative for regressions. Using this further cleaned stress data frame, we created a heatmap to view the overall correlations between stress factors. The two correlations that were immediately clear were those between self-reported stress and health risk and between biosensor-measured stress and health risk. There was also a slight correlation between sleep quality and health risk level. Since the both self-reported and biosensor stress levels correlated to health risk level, we examined the differences between those two variables and saw that people tended to slightly under-report stress level compared to what was measured with a biosensor. We decided to focus more on the biosensor stress level as it would be less prone to bias versus self-reported levels; biosensor measured stress levels were the only stress factor with an r-value of .5 compared to health risk level. We did create regressions for other stress factors compared to health risk, but none seemed to produce significant results.

Overall, we were able to determine that the only strong indicator of health risk for students was their stress level. Sleep quality also served as a weak indicator of health risk. Consequently, we conclude that students and their support networks should focus on developing strategies to reduce student stress, while focusing on sleep quality may be a good starting point. Future study could be made more valuable by expanding the demographics of those included within the study, since this study focused on a narrow age band between 18 and 24. Additionally, this dataset would be stronger if it included the past medical history of its subjects.

Works Cited

* Ziya. Student Health Data. Kaggle, n.d., [https://www.kaggle.com/datasets/ziya07/student-health-data/code. Accessed 25 Nov. 2024](https://www.kaggle.com/datasets/ziya07/student-health-data/code.%20Accessed%2025%20Nov.%202024).
* Google. https://www.google.com/search?client=firefox-b-1-d&q=how+to+change+matplotlib+color+palette. Generative AI. Accessed 30 November, 2024.
* Google. https://www.google.com/search?client=firefox-b-1-d&q=seaborn+set+color+palette. Generative AI. Accessed 30 Nov. 2024.
* XPert Learning Assistant. Accessed 30 Nov. 2024.