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 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.

We also investigated how age and gender might contribute to increased health risks. We created three different types of visualizations to answer this: a box plot for age versus health risk level, a bar chart for gender distribution by health risk level, and a violin plot for the combined effect of age and gender on health risk level. Our box plot shows that the age distribution for the three risk levels is fairly consistent, with the moderate health risk group having the broadest range of ages. The median age is close across all groups, which suggests age does not drastically vary across health risk levels.

The second visualization we created for this question is a bar chart that reveals the gender distribution across different health risk levels. It is evident that more males are present in all the risk groups. This indicates that gender doesn't play an important role in

determining the health risk level, however males are more concentrated in the moderate risk category.

The third visualization we created to answer this is a violin plot to show the combined effect of age and gender on health risk levels. The plot suggests that both age and gender influence health risks, but the relationship is not extreme, as the gender groups for each health risk level overlap considerably. For instance, both males and females in the moderate risk group show a similar age distribution, although females in the high risk group tend to be slightly younger.

Together, these visualizations suggest that while age and gender have some correlation with health risk levels, other factors may contribute significantly to determining the risk. Gender differences are particularly noticeable in the moderate risk category, but age does not show drastic variation across health levels.

The third thing we examined was how the amount of time spent studying affected health risk. Our first visualization for this was a strip plot which compared mood to the hours of study each student was performing. From this we could see that students across all moods (happy, stressed, or neutral) appear to study for a similar range of hours. Students who study excessively (above 50 hours) are distributed across all mood categories, suggesting other factors (like stress tolerance, work-life balance, or coping strategies) may influence mood. Study hours were distributed across the different mood groups, with apparent outliers appearing near the 60-hour mark for each group.

Our second visualization was a set of scatter plots with regression lines showing how study hours affect self-reported and biosensor measured stress. The regression line in each is almost flat, indicating no strong relationship between study hours and biosensor-measured stress levels. Data points are widely scattered, suggesting variability in stress levels independent of study hours. The data also shows significant dispersion, reinforcing the lack of a clear trend. Therefore, study hours alone are not a strong predictor of stress levels.

Our final visualization for this question is another strip plot which shows that there doesn't appear to be a clear or strong relationship between hours of study and health risk levels. Students with moderate, high, or low health risks study for a wide range of hours, suggesting that factors other than study hours might play a larger role in determining health risk. Study Hours have a weak negative correlation with Health Risk Numeric (-0.03), indicating a minimal relationship between the two.

The fourth question we wanted to answer was: How is sleep quality related to physical activity and stress level and /or health risk? A first look at the data showed that among 1,000 total surveyed students, 47.3% of students had good sleep quality, whereas 31.75% had moderate sleep quality while 21% of students had poor sleep quality. 207 students had good physical activity, 491 had moderate and 302 had poor physical activity

Our first visualization was a comparison of sleep quality and physical activity of students as a bar diagram. The bar plot shows that irrespective of sleep quality, the majority of students had good physical activity level, with the moderate and poor

categories being smaller. This suggests that sleep quality was not impacted by physical activity and vice versa among the students.

Next, we compared sleep quality with students' mood and visualized this in a bar chart as well. The majority of the students were reported to have a neutral mood, with a stressed mood being the least reported. When comparing among different sleep categories, the majority of students in each sleep quality group had either good or neutral moods compared to a stressed mood. This further suggests that sleep quality does not directly impact student moods.

We also examined sleep quality and its impact on stress level and overall health risk level of students via bar plot. The stress levels measured by biosensor were measured on a scale of 0-10, 0 being low and 10 being high. The bar plot we created for this comparison shows that students with all stress levels were almost uniformly distributed in all sleep quality categories. When sleep
quality was compared with overall health risk, we observed that most students had a moderate health risk level across all sleep quality categories. We noted that that the group of students with higher health risks also had a higher proportion of students with poor sleep quality than the other sleep quality categories. This implies that poor sleep quality may be directly related to increased health risk level among students.

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.
- 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.