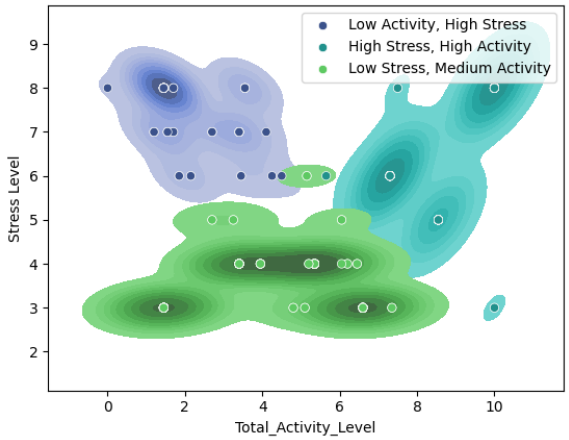
Stress vs. Activity Analysis

This project explores the relationship between total activity levels and self-reported stress across different occupations. The goal was to identify whether higher activity is associated with increased stress and whether occupation plays a meaningful role in that relationship.



The analysis includes:
- Data cleaning and occupation-level filtering
- Exploratory data analysis (EDA)
- Correlation analysis
- K-means clustering to identify patterns in activity and stress

Tools:

Python, Pandas, Seaborn, Scikit-learn, Jupyter Notebook

Findings

When looking closely at the two-cluster data, I found that the low-activity population had a negative correlation between stress and activity levels while the high-activity population had a positive correlation. I hypothesized that this correlation could have been linked to occupation, but when I looked closer at the occupation data, I found that there wasn't enough data in my dataset to get concrete conclusions from it. Even the individual occupations which showed a positive stress vs. activity correlation didn't show as strong a correlation once I looked at the raw data. This was likely due to several occupations having insufficient sample sizes for detailed analysis.

These results suggest that the relationship between activity and stress is more visible at the aggregate level than within specific occupations.

Methodology

In this project I created a combined metric called Total Physical Activity which combined the data from the Daily Steps a Physical Activity Column. This allowed me to create more readable charts and allowed be to quickly determine the correlation between different samples within clusters.

Experiments

On looking at the initial Stress vs. Physical Activity data, I saw what looked like 2–3 distinct clusters in the raw data, so I ran the k-means algorithm on the data twice, once for two clusters and once for three. I found that both cluster groups were potentially useful. The two-cluster data were more stable over time; therefore, I think it's a more accurate description of the underling patterns.

Final Observations

- Increasing total activity may reduce stress for individuals with below-average physical activity.
- For some individuals with above-average activity, lower activity levels are associated with reduced stress.
- The dataset's optimal average daily step count is ~6,200 (derived from the low-stress, medium-activity cluster).
- Occupation-level analysis was inconclusive due to limited data per occupation.

Conclusion

Distinct clusters were observed at the aggregate level, demonstrating that population structure exists in activity–stress behavior.
The two-cluster solution was selected for the main analysis because it provides a clear and interpretable separation.
However, the three-cluster solution can be useful for identifying subgroups, such as low-stress individuals.

Occupation-level analyses revealed that within-group correlations were weak and often influenced by extreme observations, highlighting that the aggregate patterns are driven primarily by differences between groups rather than consistent relationships within occupations.

These insights provide a robust and reproducible foundation for exploring population structure in activity–stress data and for guiding future work with additional or real-world datasets.

Source:

[Stress vs. Physical Activity Correlation] (https://www.kaggle.com/code/daniellevenstein/stress-vs-physical-activity-correlation)