An Investigation into The Effects of Stress Level on Sleep Health.

Word Count: 2,732

Contents

List of Figures

List of Tables

1. Introduction

Sleep is one of the most important foundations for good physical and mental health. The national sleep foundation claims that 7 – 8 hours of sleep is optimal for adults whereas young adults should be getting 7 – 9 hours of sleep a night (Espie, 2022). Despite this, many academics claim that since the COVID pandemic, sleep disorder has been on the rise. A survey by the NHS in 2022 found that 64% of young people had a problem with getting to sleep 3 or more times a week (NHS, 2022). Furthermore, an article by Gordon, Yao, Brickner and Lo (2022) found a significant rise in older adults struggling with sleep, with 30% of adults getting less than the minimum recommended 7 hours of sleep.

As studies prove that sleep health is slowly deteriorating amongst different generations, it is undoubtedly important to help gain a better understanding as to reasons why this problem happens. Analysing factors such as occupation and stress level can help determine if there is a correlation or relationship between those variables and sleep health. Understanding this can help people make informed decisions about lifestyle and career choices to provide support and advice on what impacts their sleep ability. The report therefore believes this problem is important to study and undergo further analysis.

This report will examine a dataset about sleep health to explore patterns and relationships amongst variables. Utilising a variety of statistical and machine learning techniques to determine if there is a significant corelation in sleep health and the chosen set of variables.

2. General Summary and Background of the Dataset

2.1 Source

The dataset used in this report is freely available on Kaggle, and it covers a variety of variables related to sleep patterns, daily habits and lifestyle choices. The overall dataset contains 374 observations (rows) and 13 variables (columns). However, we have narrowed down to several variables suitable for the purpose of our report. The variables we specifically focus on for this report includes occupation, quality of sleep, sleep duration and stress level. This report will be using the term *'sleep health'* to categorise the variables for quality of sleep and sleep duration.

2.2 Justification for the Dataset

The dataset initially stood out as it included a vast amount of data relating to sleeping habits. With a broad number of other variables available, it became apparent that there was an opportunity to study this topic further, to make comprehensive insights on sleep health and lifestyle patterns. There was a variety of avenues of analysis this report could have undertaken due to the plethora of variables in the dataset. For instance, the investigation could have conducted analysis on the impact of BMI or physical activities upon sleep health. The report chose a more specific investigation, so a deeper understanding of the relationship between two specific variables could be determined as opposed to brief analysis on a multitude of other factors.

3. General Analysis of the Dataset

3.1 Background and Overview of the Dataset

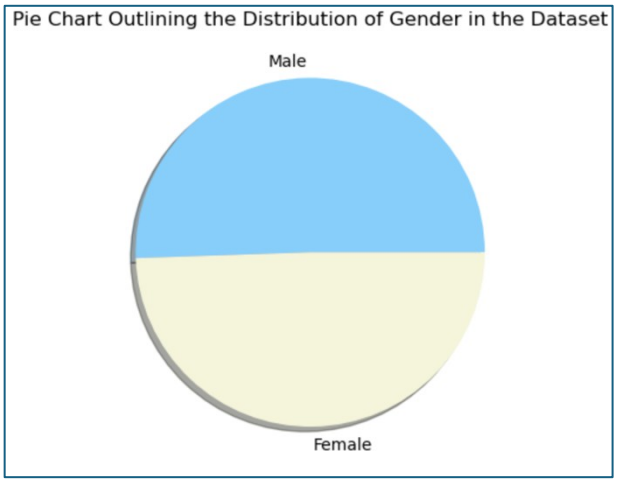

Figure 1: Pie chart showing the distribution of male and female observations in the dataset.

Figure 1 shows the spread of male and female observation across the dataset. As can be seen there is a near 50/50 split in the data with 189 observations for male and 185 for female. This means that there is no gender bias among the observations as there is equal amounts for each gender.

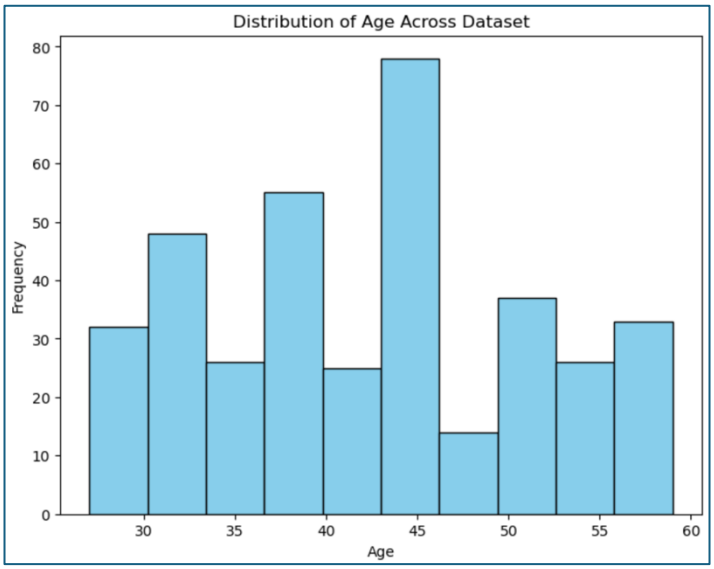

Figure 2: Histogram showing the distribution of age in the dataset.

Figure 2 shows the distribution of age amongst the dataset. As can be seen it is not normally distributed with a notable higher frequency around the 35 – 45 age range. Despite this the data still has a wide range of values allowing for non-bias from an age perspective.

3.2 Relationship Between Stress Level and Sleep Quality

With so many possible variables to choose from, a heatmap was constructed to help understand corelations between different variables. This helped solidify the choice of wanting to do stress level and sleep quality as looking at Figure 3, it can be seen there is a strong negative corelation between the two specific variables.

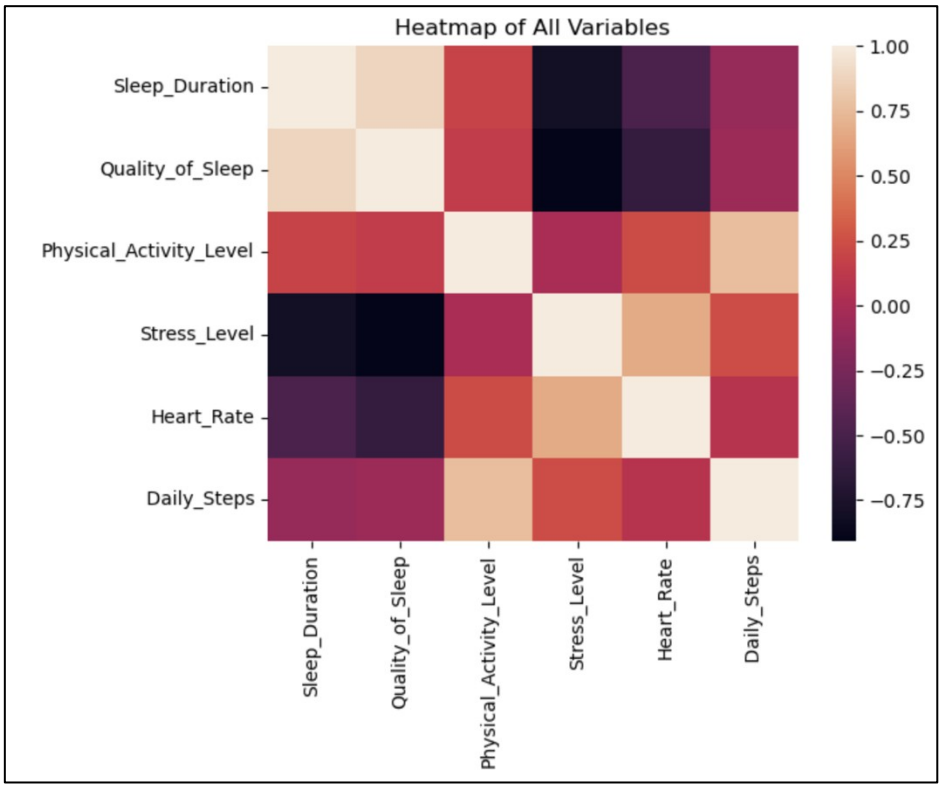

Figure 3: Heatmap of selected variables in the dataset.

Once the variables where chosen we began the analysis. The scale of values for stress level ranged from 1 – 10. However, the minimum value in the sample for stress level was 3 and the maximum was 8. Therefore, three main categories were constructed to better represent the data.

Stress Categories	Level
Low	3-4
Medium	5-6
High	7-8

Table 1: Stress Categories

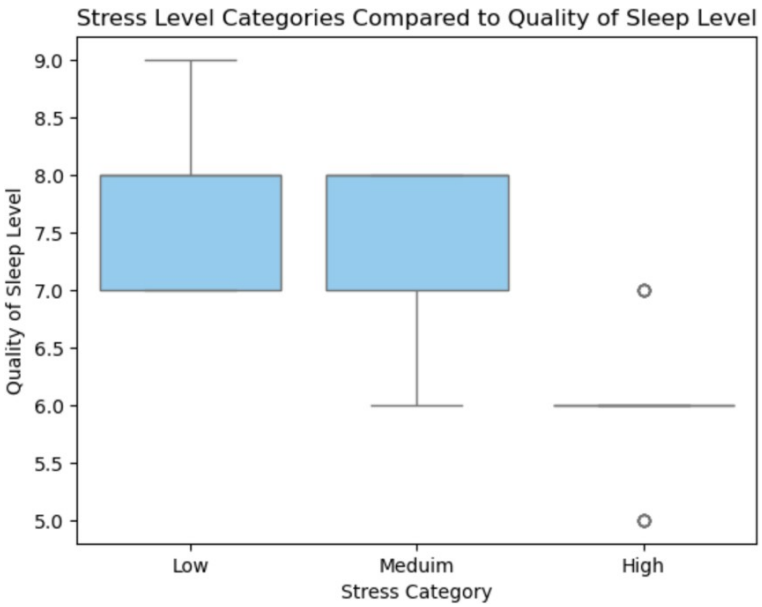

Figure 4: Boxplot of the variables stress level and quality of sleep.

As can be seen from Figure 4, the range of values for low stress level corresponds with a higher quality of sleep with the whiskers of the boxplot reaching as high as 9. The interquartile range (IQR) is between 7 and 8, meaning from initial glance, people with a low level of stress in their life have better sleep. The boxplot related to medium shows the same IQR as before however with a lower minimum value of 6. Referring to high stress level, there is a lot less data for this category displayed by the thin box. Nonetheless, with the median value of 6 as well as the outliers ranging in a lower area of the scale, point to the conclusion that people with a high level of stress receive a worse quality of sleep. Although this analysis has only come from a boxplot, so no definitive conclusions can be made without further analysis within this report.

3.3 Relating This to Occupation

A report by Kim and Lee (2015) studied how someone's job impacts their sleep ability. The report analysed 50,000 employees in Korea to help understand if there where patterns in working hours and sleep disturbances. In the explanation of their analysis, they claim that "long working hours affected sleep disturbances among male workers and female non manual workers" with a possible factor being "higher psychological work stress" (Kim and Lee, 2015).

As the chosen dataset contained occupations of the sample, this report is further able to analyse this and discover the different ranging occupations that existed.

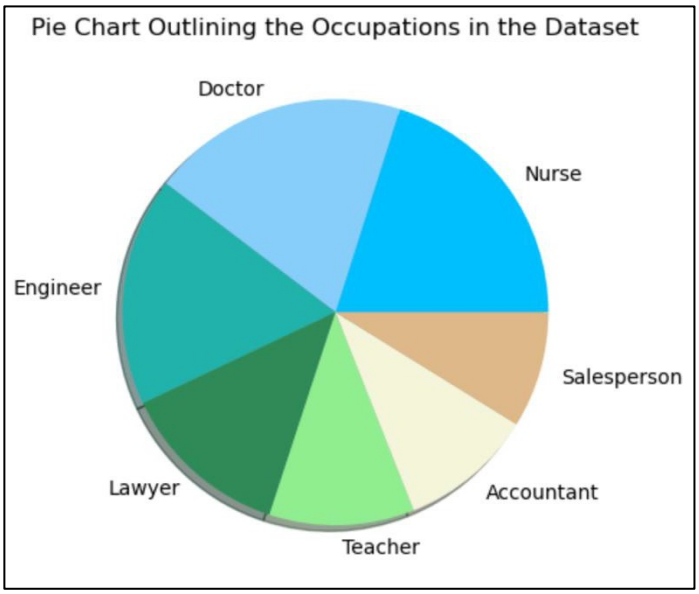

Figure 5: Pie chart outlining the number of different occupations in the data set excluding the small sample size related jobs.

Figure 5 shows a pie chart detailing the range of jobs in the sample size. All these jobs contained a sample size of 30 and above and thus were used in the analysis. However, several occupations contained a sample size of less than 30, and $N < 30$ generally do not produce a normal distribution (Uttley, 2019). Hence the following occupations were removed: scientist, software engineer, sales representative and manager.

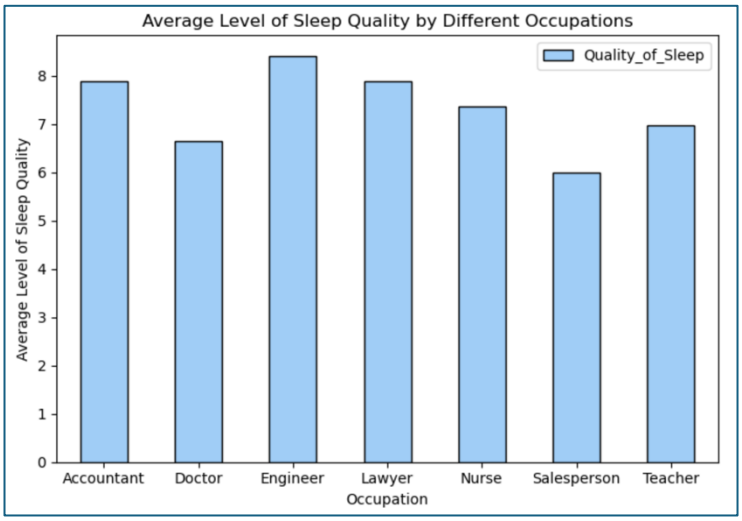

Figure 6: Bar chart outlining the quality of sleep level by different occupations.

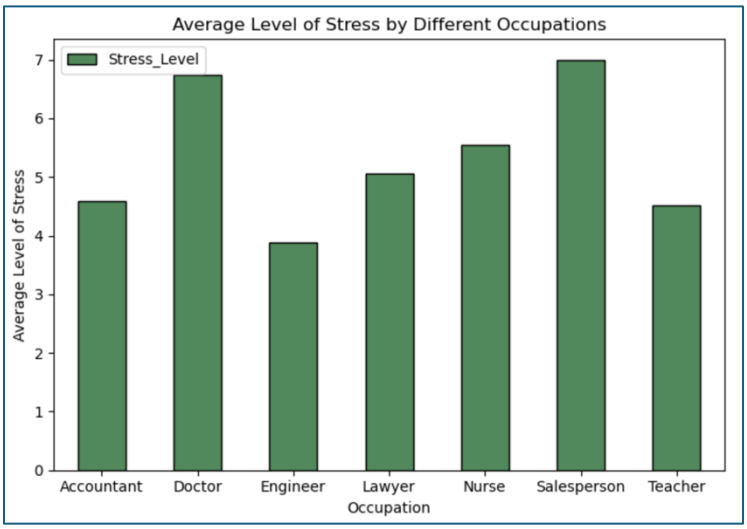

Figure 7: Bar chart outlining the average level of stress by different occupations.

Figures 6 and 7 show the average level of sleep quality and stress level amongst different occupations respectively. Figure 6 shows that occupations such as doctor and salesperson receive a lower level of sleep quality whereas jobs such as engineer, and accountant receive a much higher level of sleep quality.

Moreover, figure 7 shows the same professions; doctor and salesperson to have a higher level of stress than engineers and accountants who have a lower level of stress. From looking at these two graphs it could be presumed that there is a negative corelation between the two sets of variables, relating to the previously mentioned report by Kim and Lee (2015). However, this is only speculatory, linear regression will need to be utilised to understand the relationship further.

4. Supervised Method – Linear Regression

4.1 Introduction to Linear Regression

Supervised methods are useful for this model in classifying predictors and target variables. As lifestyle factors influence sleep quality and fluctuate daily, it is fundamental to predict sleep performance in response to changes in sleep and lifestyle habits. Linear regression helps identify relationships and make predictions into how lifestyle parameters affect sleep.

Linear regression is suitable for working on a continuous scale of values, and since the variables in this dataset are also continuous, this is an appropriate modelling approach. As mentioned, this report is focusing on the relationship between sleep health and stress; therefore, the variables sleep duration and sleep quality are selected against stress for this supervised approach. It would be interesting to predict how stress levels affect both parameters, which consequently affects an individual's sleep health. Providing comprehensive insights into how lifestyle factors interact with sleep health.

A line of best fit is created when two variables are fitted into the model, where X and y represent the independent variable and dependent variable. By fitting the model into a linear regression, the equation $y = t0 + t1 \times X$ was used. Furthermore, a score value (R^2) is calculated to assess how well the model captures the relationship between the two variables. The closer it is to 1, the more variability the model explains.

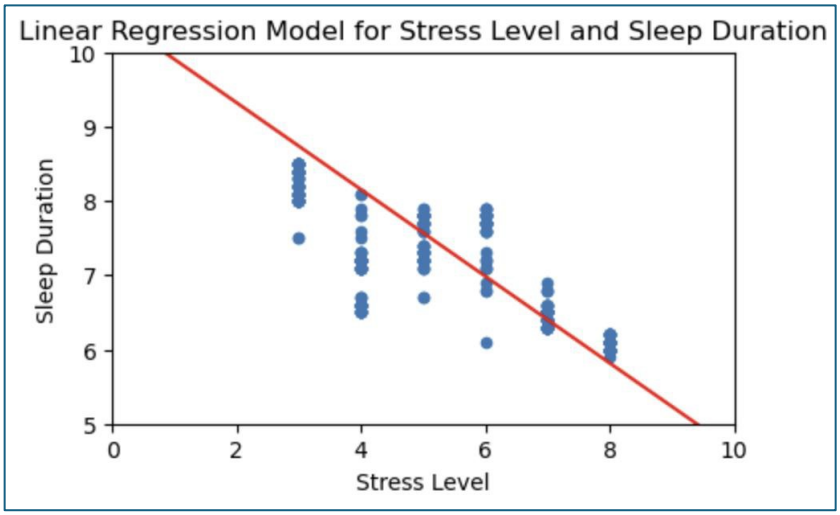

Figure 8: Linear Regression Model for Stress Level and Sleep Duration.

4.2 Linear Regression for Stress Level and Sleep Duration

For the first model, the independent variable is the stress level, and the dependent variable is the sleep duration. A score value of 0.651 was obtained. Seen in Figure 8, most points surrounding the line of best fit are relatively scattered, suggesting a moderate variability. With this predicted model, a coefficient value of -0.36 and intercept value of 9.07 was calculated. The negative coefficient suggests a negative relationship between stress and sleep duration; for every additional hour of sleep, stress levels are predicted to decrease by 0.36 levels. For example, a predicted stress level of 8 is associated with an estimated sleep duration of 6.2 hours, which is below the recommended daily amount of sleep for good sleep health (Espie, 2022). This aligns with the existing literature where higher levels of perceived stress are associated with a poorer sleep performance (Charles et al., 2011).

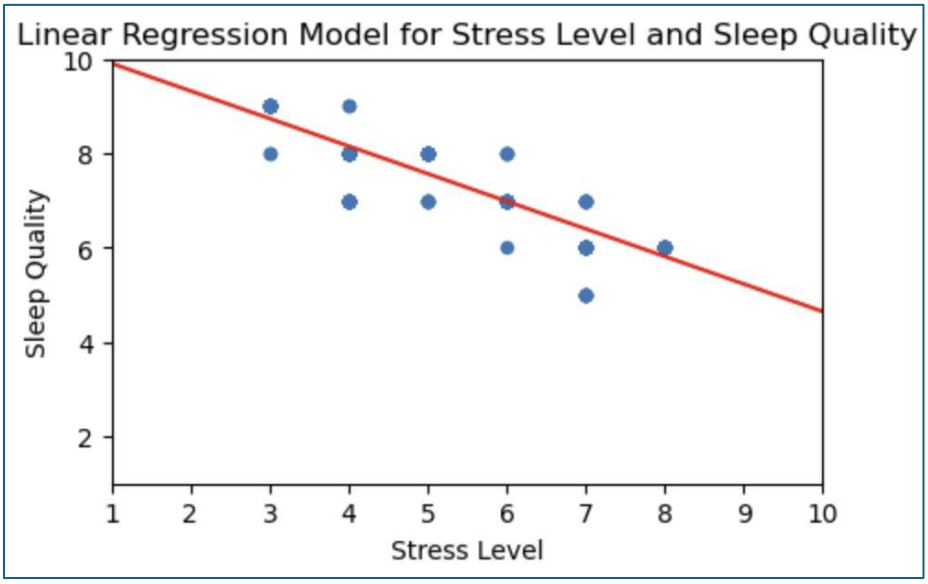

Figure 9: Linear Regression Model for Stress Level and Sleep Quality.

4.3 Linear Regression for Stress Level and Sleep Quality

Furthermore, a similar model was created by looking between stress level and quality of sleep. A coefficient value of -0.584 and intercept value of 10.491 was calculated. Again, a negative coefficient is obtained, indicating another negative relationship between stress and sleep health, specifically regarding sleep quality. These numbers suggests that higher stress levels are associated with lower sleep quality, further supporting the negative impact of stress on overall

sleep health. For example, a high predicted stress level of 9 is associated with an estimated sleep quality of 5.232 rating out of 10, which is considered a low quality of sleep. However, this model had a score value of 0.823, which has a better variability than the previous model. Figure 9 shows the points are less scattered around the line of best fit.

Both linear regression models offers valuable insights into the negative impact of stress on sleep health. This encourages various occupations such as doctor and salesperson, that on average have a higher stress level, to effectively manage their stress to enhance sleep health.

Linear Regression Model	**Score Value (R^2)**	**Coefficient Value**	**Intercept Value**
Stress and Sleep Duration	0.651	-0.359	9.074
Stress and Quality of Sleep	0.823	-0.584	10.491

Table 2: Results from the Linear Regression Models.

5. Unsupervised Analysis

5.1 Introduction to Unsupervised Analysis

Unsupervised analysis allows the exploration of data without prior assumptions (Xie, Girshick and Farhadi, 2016; Igual and Seguí, 2017; Sheng and Li, 2021). It helps uncover patterns for improving sleep health based on individual lifestyles. Using this method is useful as this dataset lacks the predefined labels for “poor” or “healthy” sleep habits (Erman, Korosec and Suklan, 2015). Gupta, et al. (2021) states that as class labels are not predefined in the cluster analysis, it is important to remove non-essential information to improve clustering results. As the report focuses on stress level and sleep health, 3 variables were considered: stress level, quality of sleep, and sleep duration.

5.2 Hierarchical Agglomerative Clustering (HAC)

HAC is considered the bottom-up approach, it begins with each data point, and then merges with other datapoints with similarities (Erman, Korosec and Suklan, 2015). In the context of sleep health, this method can assist in identifying underlying patterns and relationships between different factors that influence sleep quality. By clustering similar individuals based on their sleep-related characteristics, valuable insights are gained that can inform targeted interventions and improve overall sleep health.

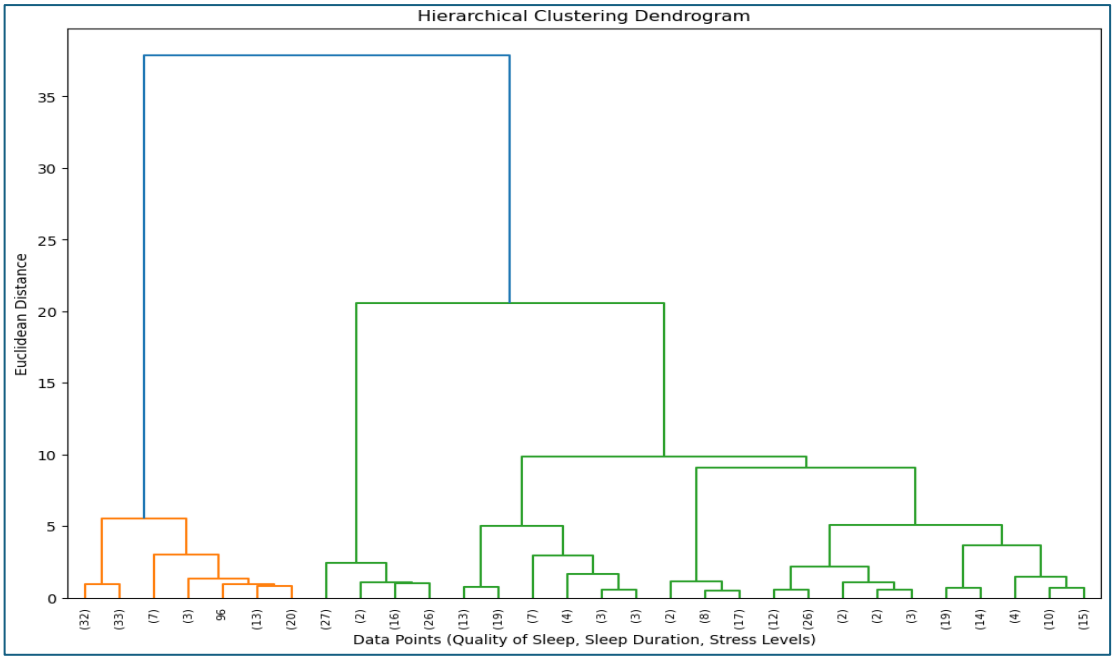

Figure 10: Hierarchal Agglomerative Clustering

5.3 HAC vs K-Means:

K-Means is useful when there is a predetermined number of clusters known beforehand (Likas, Vlassis and J. Verbeek, 2003). Consequently, as this was unknown, HAC was used. Tate (2023) argues that it is important to understand the hierarchical structure, when doing exploratory analysis, as it could offer significant insights that K-Means might overlook.

Figure 11 displays only 14 distinct data points because of overlapping values. This overlap occurs because both stress level and quality of sleep are measured on a limited scale ranging from 1 to 10. Figure 10 provides a more detailed outlook of the clusters, revealing distinct patterns. Additionally, it effectively separates data points, allowing for a comprehensive view of how stress levels and sleep quality vary among individuals.

Figure 11: K-Means Clustering Quality of Sleep & Stress Level.

5.4 Analysis of HAC Graph

From Figure 10, the dendrogram has separated the data points into 3 distinct clusters. This is useful as we can identify patterns associated with overall sleep health. Cluster 1 (orange) represents individuals with lower sleep duration, lower sleep quality, and higher level of stress. Meanwhile cluster 2 (green) suggests individuals with moderate sleep duration, sleep quality and stress. This group could be considered the most balanced. Cluster 3 (blue) highlights the individuals with the longest sleep duration, highest sleep quality, and lowest stress levels. The large Euclidean distance between the orange and green clusters highlight that they are different from each other in terms of sleep quality, duration, and stress levels (Krislock and Wolkowicz, 2012).

Based on the analysis gained from the clustering, the green has majority of the clusters. Consequently, it confirms that the individuals in the dataset have average levels of overall sleep health and stress. These individuals could improve their overall health from stress management techniques, more regular physical activity, and sleep hygiene practice.

6. Reflections

We believe the dataset used in this report was worthwhile and productive. The data was easily accessible, easy to interpret and contained a plethora of variables, many of which we never touched upon. Furthermore, the scale used for certain variables i.e. stress level and sleep quality was consistent throughout, allowing regression and other types of analysis to be used efficiently. The sample size was an equal split of male and female observations, meaning there was no gender bias, as well as a large variance of ages allowing many different perspectives to be obtained.

However, the dataset did come with some drawbacks. The sample size was notably small with only 374 observations per variable. When analysing specific variables such as stress level in a box plot, the results looked partly bizarre with a lack of observations for the high stress category. In addition, certain occupations having too small a sample size, shrinking the possible area of analysis. Furthermore, linear regression does not capture other confounding variables, such as age or underlying health conditions. Correlation does not mean causation, suggesting that stress levels may not solely impact sleep durations. Therefore, this supervised method only provides a rough prediction to our analysis rather than a comprehensive explanation, as all possible parameters that influence sleep health are not captured.

7. Conclusion

In conclusion, our analysis offered a fair baseline into investigating sleep health with various parameters. From the regression models, the report found that there was a negative correlation between stress level and sleep health, meaning for example that when stress level increases, sleep health overall decreases. The supervised method provided a great insight into how stress negatively impacts sleep health. It supports the existing literature and provides stronger evidence for occupations with high stress levels to be more aware of the potential impact on sleep health. The hierarchal agglomerative method analysed the overall sleep health and discovered patterns and relationships. Ultimately, it showed most of the clusters are related to the colour green, showing most individuals in the sample has overall average sleeping health.

Appendix

Software versions and packages

Python: Version 3.12.7

Data from: https://www.kaggle.com/datasets/uom190346a/sleep-health-and-lifestyle-dataset

Packages used:

- numpy
- pandas
- matplotlib.pyplot
- seaborn
- sklearn.linear_model
- StandardScaler from sklearn.preprocessing
- datasets from sklearn
- metrics from sklearn
- cluster from skleaern
- scale from sklearn.processing
- KMeans from sklearn.cluster
- hierarchy from scipy.cluster
- dendrogram, linkage from scipy.cluster.hierarchy

References

Aastha Gupta, Himanshu Sharma, and Anas Akhtar (2021) 'A COMPARATIVE ANALYSIS OF K-MEANS AND HIERARCHICAL CLUSTERING', *EPRA International Journal of Multidisciplinary Research (IJMR)*, pp. 412–418. Available at: https://doi.org/10.36713/epra8308.

Charles, L.E. *et al.* (2015) 'Association of Perceived Stress with Sleep Duration and Sleep Quality in Police Officers'.

Erman, N., Korosec, A. and Suklan, J. (2015) 'PERFORMANCE OF SELECTED AGGLOMERATIVE HIERARCHICAL CLUSTERING METHODS', *Innovative Issues and Approaches in Social Sciences*, 8, pp. 180–204. Available at: https://doi.org/10.12959/issn.1855-0541.IIASS-2015-no1-art11.

Espie, C.A. (2022) 'The "5 principles" of good sleep health', *Journal of Sleep Research*, 31(3), p. e13502. Available at: https://doi.org/10.1111/jsr.13502.

Gordon, N.P. *et al.* (2022) 'Prevalence of sleep-related problems and risks in a community-dwelling older adult population: a cross-sectional survey-based study', *BMC Public Health*, 22, p. 2045. Available at: https://doi.org/10.1186/s12889-022-14443-8.

Hirshkowitz, M. *et al.* (2015) 'National Sleep Foundation's sleep time duration recommendations: methodology and results summary', *Sleep Health*, 1(1), pp. 40–43. Available at: https://doi.org/10.1016/j.sleh.2014.12.010.

Igual, L. and Seguí, S. (2017) 'Unsupervised Learning', in L. Igual and S. Seguí (eds) *Introduction to Data Science: A Python Approach to Concepts, Techniques and Applications*. Cham: Springer International Publishing, pp. 115–139. Available at: https://doi.org/10.1007/978-3-319-50017-1_7.

Kim, J. and Hyung, W. (2013) 'Dynamical model for gamification of learning (DMGL)', *Multimedia Tools and Applications*, 74. Available at: https://doi.org/10.1007/s11042-013-1612-8.

Krislock, N. and Wolkowicz, H. (2012) 'Euclidean Distance Matrices and Applications', in M.F. Anjos and J.B. Lasserre (eds) *Handbook on Semidefinite, Conic and Polynomial Optimization*. New York, NY: Springer US, pp. 879–914. Available at: https://doi.org/10.1007/978-1-4614-0769-0_30.

Likas, A., Vlassis, N. and J. Verbeek, J. (2003) 'The global *k*-means clustering algorithm', *Pattern Recognition*, 36(2), pp. 451–461. Available at: https://doi.org/10.1016/S0031-3203(02)00060-2.

Murtagh, F. and Legendre, P. (2014) 'Ward's Hierarchical Agglomerative Clustering Method: Which Algorithms Implement Ward's Criterion?', *Journal of Classification*, 31(3), pp. 274–295. Available at: https://doi.org/10.1007/s00357-014-9161-z.

NHS England Digital (2022) *Part 2: Sleep, loneliness, activities and health behaviours*. Available at: https://digital.nhs.uk/data-and-information/publications/statistical/mental-

health-of-children-and-young-people-in-england/2023-wave-4-follow-up/part-2-sleep-loneliness-activities-and-health-behaviours (Accessed: 5 November 2024).

Sheng, J. and Li, W.V. (2021) 'Selecting gene features for unsupervised analysis of single-cell gene expression data', *Briefings in Bioinformatics*, 22(6), p. bbab295. Available at: https://doi.org/10.1093/bib/bbab295.

Tate (2023) *Comparing DBSCAN, k-means, and Hierarchical Clustering: When and Why To Choose Density-Based Methods*, *Hex*. Available at: https://hex.techcomparing-density-based-methods (Accessed: 5 November 2024).

Uttley, J. (2019) 'Power Analysis, Sample Size, and Assessment of Statistical Assumptions—Improving the Evidential Value of Lighting Research', *LEUKOS*, 15(2–3), pp. 143–162. Available at: https://doi.org/10.1080/15502724.2018.1533851.

Vijaya, Sharma, S. and Batra, N. (2019) 'Comparative Study of Single Linkage, Complete Linkage, and Ward Method of Agglomerative Clustering', in *2019 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COMITCon)*. *2019 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COMITCon)*, pp. 568–573. Available at: https://doi.org/10.1109/COMITCon.2019.8862232.

Xie, J., Girshick, R. and Farhadi, A. (2016) 'Unsupervised Deep Embedding for Clustering Analysis', in *Proceedings of The 33rd International Conference on Machine Learning*. *International Conference on Machine Learning*, PMLR, pp. 478–487. Available at: https://proceedings.mlr.press/v48/xieb16.html (Accessed: 5 November 2024).