A Sleep Analysis on Factors that Correlate with Longer Sleep Duration

Zoha Hassan
Department of Computer Science
Loyola University Chicago
Chicago, USA
zhassan@luc.edu

Fiona Nicdao
Department of Computer Science
Loyola University Chicago
Chicago, USA
fnicdao@luc.edu

***Abstract*—In this project, we investigated which factors, such as caffeine intake, occupation, stress levels, physical activity, BMI categories, and sleep disorders impact human sleep quality the most to find factors that correlate with longer sleep duration. We cleaned and prepared four datasets related to sleep, and analyzed them using techniques including bar plots, histograms, box plots, outlier detection methods, PCA, tSNE, and UMAP. Our findings were that regardless of factors such as occupation and sleep disorders, high physical activity, normal BMI, and low stress environments improve sleep duration. This highlights the importance of holistic health as it impacts sleep duration and quality.**

I. Introduction

There is a lot of research on the topic of sleep, especially regarding how to improve human's sleep quality and duration since sleep has an extremely large impact on health and wellness. As graduate students, sleep is usually at a minimum and we understand the importance of sleep to one's health and thought it would be an important study to see what factors affect one's sleep and what can be done to improve one's quality of sleep.

II. Methods

We decided to use available data to find out: Which factors improve human sleep quality the most? We explored various factors such as caffeine intake, occupation, stress levels, physical activity levels, and more. To address this question, we used bar plots, histograms, box plots, outlier detection methods, PCA, tSNE, and UMAP to find connections within and between our four datasets after understanding them.

III. Data

We worked on four different dataset that related to how sleep affects cognitive performance, productivity, health, and sleep efficiency. The datasets are procured from Kaggle and are linked below.

- Sleep Deprivation & Cognitive Performance
- Sleep Cycle & Productivity
- Sleep Health & Lifestyle Dataset
- Sleep Efficiency Dataset

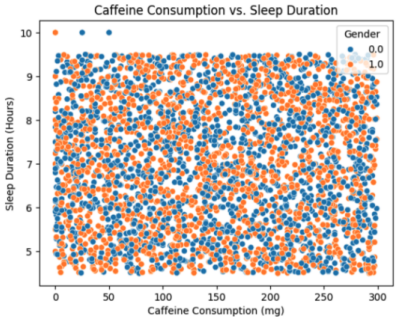
IV. Results

A. Age effects on Sleep

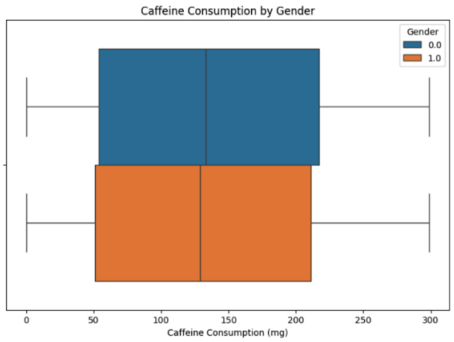
In the first section, we analyze all the datasets and group everyone by age groups, gender, and amount of sleep in hours per day. We noticed that a majority of our samples are aged 25 - 55 years old and only 15 samples in the 65 -100 years old group. Then looking at the distribution of gender,, we see that there is an even amount of male and female with around 2000 samples and about 1600 samples that identify as other. Then, looking at the distribution of sleep for each age group, we notice that regardless of the age group, a majority of people fall between 4.5 and 9.5 hours of sleep, and very few people land outside of that range. Also in the 65 - 100 age group nothing can be concluded because of the small sample size of 15 people.

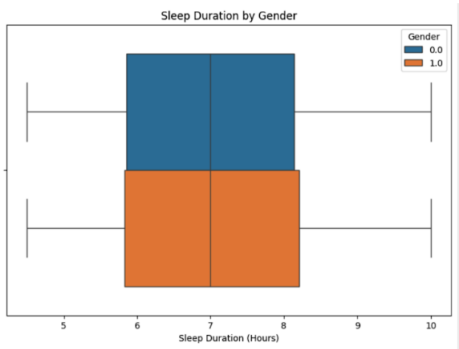
B. Caffeine and Gender effects on Sleep

One of the factors we were interested in seeing its effect on sleep duration was caffeine, and we decided to specifically see if caffeine had different impacts by gender. For this section, we chose to analyze the datasets Sleep Efficiency and Sleep Cycle Productivity. We began by creating a new dataset, df-caffeine, by merging the attributes age, gender, sleep hours, and caffeine intake from the two datasets. In these two datasets, male and female were the only genders listed so we converted the categorical values to numeric. After our dataset was created, we created a scatterplot to compare caffeine consumption vs sleep duration by gender, however this resulted in an unclear scatterplot that would be impossible to extract correlations from as there were too many data points to make sense of. After this, we tried to narrow down the group to observe, and decided to explore the impacts of caffeine intake on females ages 18-25. Despite this smaller sample size, however, there were still no correlations that could be extracted from this scatterplot.



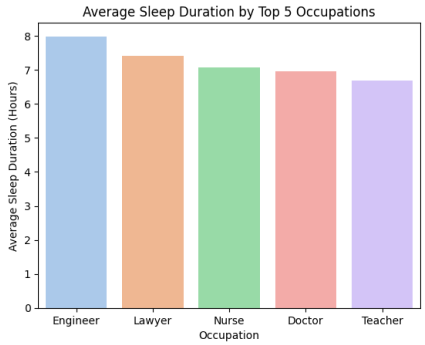
Continuing our exploration of this factor, we observed that it seems males and females consume roughly similar amounts of caffeine regardless of age group, and also experience roughly the same amounts of sleep duration. We then conclude that gender is not a good separator for caffeine and duration of sleep.



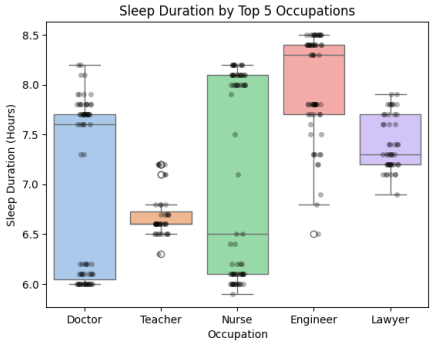


C. Occupation effects on Sleep Duration

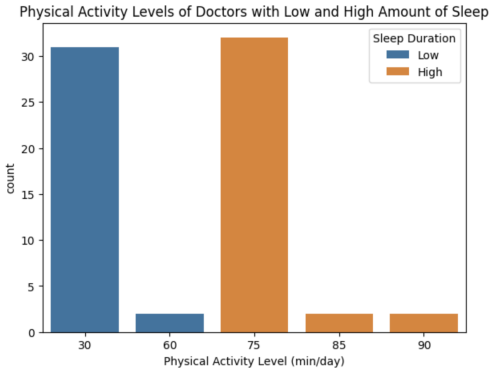
The next factor we wanted to explore was the impact of various occupations on sleep duration. For this, we used the Sleep Health and Lifestyle dataset. To begin, we listed the top 5 occupations as well as all the occupations in the dataset and their counts. We found that the top 5 occupations were nurse, doctor, engineer, lawyer, and teacher. Exploring the average sleep durations of each of these occupations, we found that engineers have the highest average sleep duration, then lawyers, nurses, doctors, and lastly teachers.

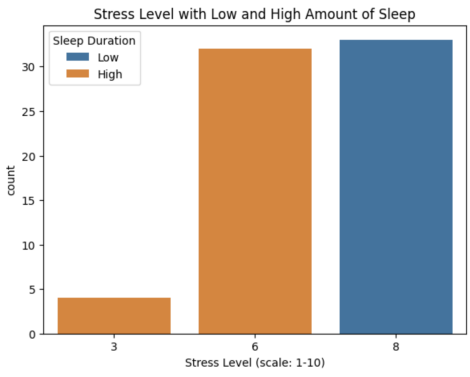


With further exploration of this using a boxplot and stripplot, we found that teachers have the smallest range of sleep duration, from around 6.6-6.7 hours of sleep, and nurses have the biggest range of sleep, from around 6.2-8.1 hours of sleep. Engineers have the highest median of sleep, around 8.3 hours.



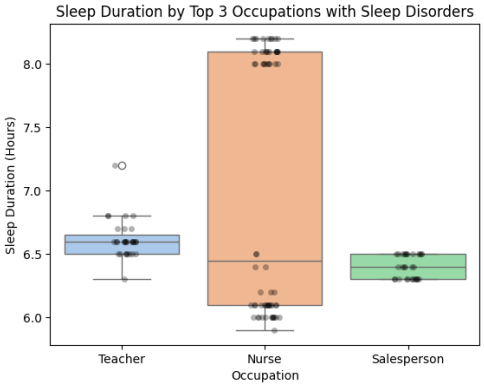
From the chart on sleep duration by the top five occupations, we notice that there are two clear groups of doctors with low amounts of sleep (less than 6.5 hours) and doctors with high amounts of sleep (greater than 7.5 hours). Therefore, we examined these two groups and compared their attributes and found two metrics that the two groups diverged, namely, physical activity level and stress level. Physical activity level measures the amount of exercise the sample partakes in minutes per day and stress level is a measure of stress on a scale of 1 to 10. From our evaluation we observed that doctors with high amounts of sleep tend to have higher levels of physical activity and report lower levels of stress than doctors with low amounts of sleep.



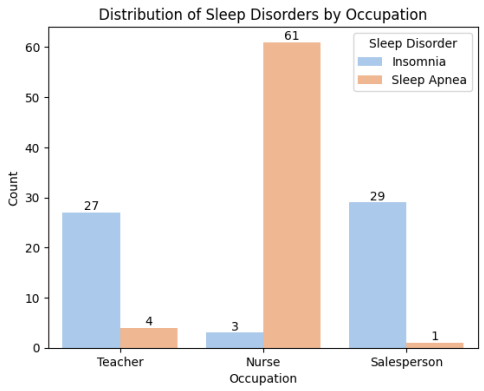


1) Which occupation experiences the most sleep disorders?: As a natural progression, the next question we wanted to answer was which occupations experience the most sleeping disorders. For this, we began by creating a new dataframe called df_sleep_disorder to isolate all the samples in df_health that listed a sleeping disorder (not "None"). Using that, we listed the top occupations with sleeping disorders, and decided

to explore the top 3: nurses, teachers, and salespersons since they had the highest counts, and the fourth and fifth highest were small counts. Similar to the analysis for sleep duration for the top 5 occupations above, we created a boxplot and stripplot the sleep duration for the top 3 occupations with sleeping disorders. We found that teachers have a median of 6.6 hours and salespersons have a median of 6.3 hours, and both of these occupations have the smallest ranges of sleep within the three with teachers range being 6.5-6.7 hours and salespersons range being 6.3-6.6 hours. For nurses, the median sleep duration is 6.4 hours, but they have a very large range from 6.1-8.1 hours of sleep.

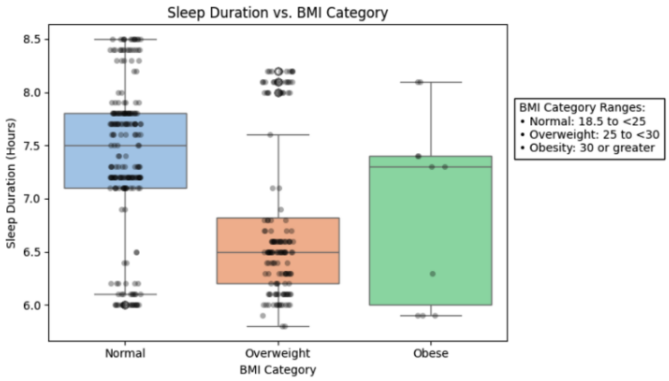


We also wanted to explore what the distribution of sleep disorder within these three occupations was. The sleep disorders within our dataset were sleep apnea and insomnia. With a countplot, we were able to observe that teachers and salespersons experience insomnia more than sleep apnea. Nurses experience more sleep apnea than insomnia. However, it is important to note that we are simply looking at a small sample size which could be contributing to these results.

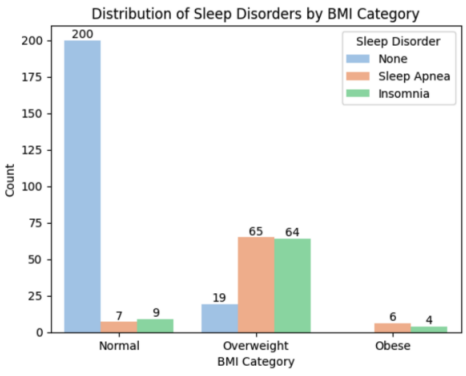


D. Exploring relationships between BMI Category, Sleep Duration and Sleep Disorders.

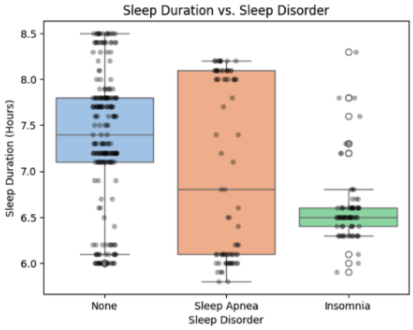
Another factor whose impact we wanted to see on sleep duration was BMI Category, as well as how each sleep disorder impacts sleep duration, and the relationship between BMI Category and sleep disorders. Continuing to use df-health, we found that the BMI categories listed in this dataset were Overweight, Normal, Obese, and Normal Weight. We combined the categories Normal and Normal Weight, since they meant the same thing. Then we listed out the counts of the samples of each category, and found there were 216 normal samples, 148 overweight samples, and 10 obese samples. For some context, the BMI index for each category is as follows: Normal 18.5 to less than 25, Overweight 25 to less than 30, Obese 30 or greater. Using a box-plot to display the relationship between BMI category and sleep duration, we found that those with a normal BMI have the highest sleep duration median, at 7.5 hours, and then obese at 7.4 hours. The overweight BMI has the smallest median of 6.5 hours. For the overweight BMI, the range is 6.3-6.8 hours of sleep. However, this group also has many outliers that indicate there are a few overweight people that get around 8 or more hours of sleep. The obese BMI has the highest sleep duration range, from 6-7.5 hours of sleep. However, this category also has very little samples in the dataset, only 10 compared to "Normal" 195 and "Overweight" 148 so that needs to be taken into consideration whenever the obese category is referred to in our project.



Taking a deeper look at each BMI category, we wanted to see which sleeping disorders impact each of these groups. Another countplot showing the distribution of sleep disorders by BMI category shows that an overwhelming majority of people with a normal BMI do not have sleeping disorders (or are undiagnosed), but some do experience sleep apnea and insomnia. A majority of the overweight BMI category experiences either sleep apnea or insomnia, which a minority does not (or is undiagnosed). For the obese BMI, we cannot come to a conclusion as there were only 10 samples, however, we will note that of those 10 samples all of them reported a sleeping disorder. This could indicate that further research could be done about this group's tendencies of having sleeping disorders.

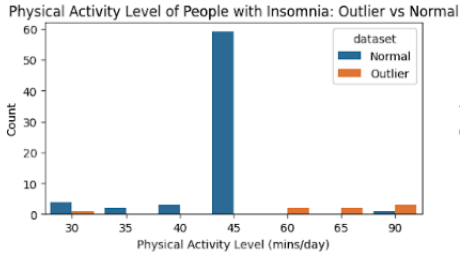


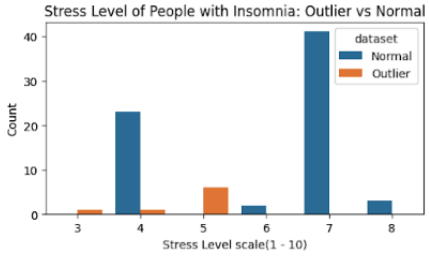
Next, we looked at sleep duration vs sleep disorders, and using a boxplot found that as expected, those with insomnia sleep much less than those with sleep apnea. The insomnia group's median sleep duration is 6.5, and the sleep duration ranges from 6.4-6.6 hours of sleep. There are many outliers with this group, however, with many people sleeping more than these values and many sleeping even less. The sleep apnea group's median sleep duration is 6.9 hours, but the range is very large from 6.1-8.1 hours of sleep. So, it seems like sleep duration might be independent of sleep apnea. Those with no sleep disorder, or are undiagnosed, range from 7.1-7.8 hours of sleep, and their median is 7.3 hours. However this group has the most outliers, with some people sleeping about 6 hours and up to 8.5 which could be explained because they may not be diagnosed.



E. Outliers

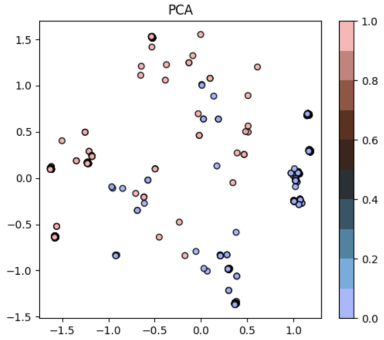
Looking at the chart of Sleep Duration vs Sleep Disorder we notice there are quite a few outliers for the insomnia boxplot, those outliers signify people who are diagnosed with insomnia and reported more sleep (greater than 7 hours) than the majority of people with insomnia. For this reason, we look at the attributes that distinguish this group of outliers. We observed that the small subset of people with insomnia and higher amounts of sleep duration have higher levels of physical activity, more daily steps, and tend to have lower heart rates. Also, we notice the outliers report lower levels of stress and higher levels of quality of sleep. In addition, the group of outliers are more likely to be normal weight, however there are some people who have a BMI in the obese category. Below are only two charts out of six factors we examined.

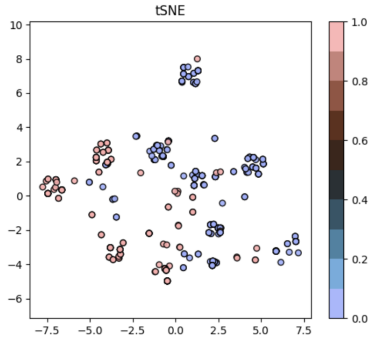




F. PCA, tSNE, UMAP for Sleep Health and Lifestyle Dataset

The hope of implementing PCA, tSNE, and UMAP is to have two clusters, one for people with sleep disorder (pink) and another for people who does NOT have a sleep disorder (blue). From the PCA graph, it is not good at separating; there is a small cluster on the right bottom side of the graph of points that do not have a sleeping disorder. From the tSNE graph, there are more split clusters of about 3 have a sleep disorder. While there are about 3-6 clusters of people who do not have a sleeping disorder. tSNE does a better job of clustering than PCA but the clusters do overlap a lot and split into multiple groups. From the UMAP (not pictured), a lot of points overlap and make tight clusters, nothing can be discerned.





V. Conclusion

Our analysis shows that, regardless of occupation or the presence of a sleep disorder, individuals who engage in longer periods of daily physical activity and report lower stress levels tend to sleep longer. Additionally, maintaining a normal BMI is also associated with increased sleep duration.