**CCT College Dublin**

**Assessment Cover Page**

|  |  |
| --- | --- |
| **Module Title:** | Strategic Thinking |
| **Assessment Title:** | CA 2 – Capstone Project Proposal |
| **Lecturer Name:** | James Garza |
| **Student Full Name:** | Arthur Speggiorin Verza |
| **Student Number:** | SBA23382 |
| **Assessment Due Date:** | 29/10/2023 |
| **Date of Submission:** | 28/10/2023 |

**Declaration**

|  |
| --- |
| By submitting this assessment, I confirm that I have read the CCT policy on Academic Misconduct and understand the implications of submitting work that is not my own or does not appropriately reference material taken from a third party or other source. I declare it to be my own work and that all material from third parties has been appropriately referenced. I further confirm that this work has not previously been submitted for assessment by myself or someone else in CCT College Dublin or any other higher education institution. |

**Introduction**

For this assignment, I was asked to find a dataset that piqued my interest and create a business plan that formulates a hypothesis and questions to uncover meaningful answers. After exploring a few intriguing topics on Kaggle, I stumbled upon a dataset about sleep health and lifestyle. The dataset provides an excellent opportunity to analyse the correlation between sleep quality and factors such as physical activity, weight, sleep duration, stress level, and any sleep disorder.

According to Jansen (2020), sleep is interdisciplinary because it touches every aspect of health and is essential for every individual. Furthermore, Ramar et al. (2021) highlight that sleep is a biological necessity, and inadequate sleep can lead to disorders that impact health and well-being. Understanding the factors affecting sleep quality is crucial for improving overall health outcomes.

The Sleep Health and Lifestyle dataset contains approximately 400 rows and 13 columns, offering a comprehensive view of sleep habits and lifestyle information. The dataset includes variables like gender, age, occupation, sleep duration, stress level, sleep quality, physical activity level, blood pressure, heart rate, daily steps, and sleep disorders. This wide range of information allows for a detailed analysis of how lifestyle factors and demographic variables influence sleep quality.

By exploring this dataset, I aim to uncover patterns and correlations that can help predict sleep quality and stress levels. My goal is to provide valuable insights that can guide interventions to improve sleep health and lifestyle habits, ultimately leading to better overall well-being.

**Objectives**

The objective of this project is to predict sleep quality and stress levels among individuals and identify the most significant effects of poor sleep quality on daily tasks. By analyzing and relating occupation with stress levels, I aim to understand the impact on daily physical activities, sleep quality, and the duration of a night's sleep. Additionally, I will explore how sleep disorders influence sleep quality. The insights gained will help shed light on the correlation between these lifestyle factors and overall well-being.

To achieve this goal, I will utilize various machine learning models learned throughout the course. The models presented in class will be evaluated to determine the best fit for the selected dataset. I will carefully choose models that can effectively analyze the relationships between sleep quality, stress levels, occupation, and other lifestyle factors.

For this project, I plan to use supervised learning. As Müller and Guido (2016, p. 25) state, "Supervised learning is used when we want to predict a certain outcome from a given input and we have examples of input/output pairs." Given the dataset available, supervised learning is well-suited to predict sleep quality and stress levels.

After examining the chosen dataset, I believe it's crucial to consider all available information to understand the true impact of sleep quality on people's lives. Factors like occupation, lifestyle behaviours, and sleep disorders must be carefully analysed to develop a comprehensive understanding of how sleep quality affects daily life. By using a holistic approach, this project aims to provide valuable insights into improving sleep health and reducing stress.

**Problem definition**

When exploring the chosen dataset, I noticed that some occupations interestingly relate to stress levels and other factors such as sleep quality, time spent on physical activity during the day, sleep duration, and whether a person has a sleep disorder. Together, these factors collectively influence a person's stress level.

With this project, I aim to predict the correlation between sleep quality, physical activity, occupation, sleep disorders, and sleep duration in relation to stress levels. By doing so, I hope to provide useful insights that could help individuals reduce their stress levels and improve their overall sleep quality.

**Scope**

In this project I am going to analyse one dataset about stress level and the correlation with sleep duration, sleep quality and physical activity. My dataset has 13 variables, but after a review of the dataset we decided to exclude some columns from the analysis.

I decided to not exclude any column from my dataset, because it may affect and have a positive or negative relation with the sleep quality. I will analyse some of the points, won’t analyse all because it will be so many variables and can be confused.

The stress level and the sleep quality are inversely related  and also looking into the other variables, like job occupation, sleep disorder. I will be looking at I can see that the level of stress changes accordingly depending on the job occupation and quality of sleep and sleep duration of a person on a night. According to my initial analysis of the dataset, people who have jobs where requires longer shifts and people with higher level of stress have worst sleep quality. Also, people with some sleep disorder tend to have worst sleep quality.

With this capstone project I will try to deliver a model where it will be possible to predict based on some variables the level of stress someone will have and then will be able to avoid high levels of stress providing support to improve the other variables that would affect it.

As the capstone project is a 2 semester project I need to have a proposed timeline so I can have a good project management plan.  I am thinking about the platform Monday.com to use as a good way of keeping track of the following phases. I have below some idea of how it will be divided by week duration, remembering that this is just an idea as we are still on the way of learning the topics to be used on the project.

* Dataset Analysis and preparation ( Approx. 3 weeks)
* Select Machine Learning Model to be used and start working on applying it  (Approx. 2 weeks)
* Apply Machine learning Model and do tests to see if will work as expected (Approx. 2 Weeks)
* Reports building will be done during the tests and selection of the model

**Data Sources**

The data selected was taken from Kaggle linked [here](https://www.kaggle.com/datasets/uom190346a/sleep-health-and-lifestyle-dataset/data), I would like to make an observation about the dataset as this data set is a synthetic one that was created only for illustrative purposes and no real people information was used.

**Ethic**

In my view as mentioned previously I did not have any issues with requesting permission to use the data set as it was publicly available and also no ethical issues due the dataset being a synthetic one and not personal information that could label a person was presented.

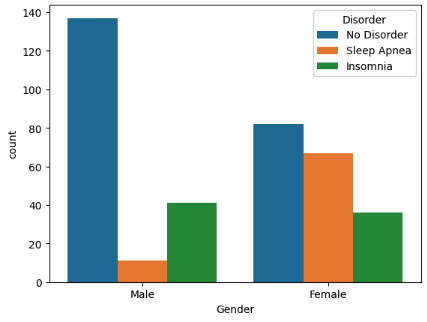
**Data processing and results**

Initially, I started by renaming the columns to shorter names to improve visualization when using commands like head. I also checked for missing values and duplicates. In the sleep\_disorder column, I found 219 missing values, which indicated that 219 individuals did not have any sleep disorder. Thus, I replaced the missing values with "No Disorder."

Analyzing sleep disorders by gender, it became evident that men have fewer sleep disorders than women, as shown in Figure 1. Further examining sleep disorders based on occupation, individuals working longer shifts or in lower-paying jobs tend to have a higher number of sleep disorders (Figure 2).

These initial findings underscore that certain demographic and occupational factors significantly impact sleep health. Understanding these correlations can guide strategies to improve sleep quality, particularly for those in high-risk groups.

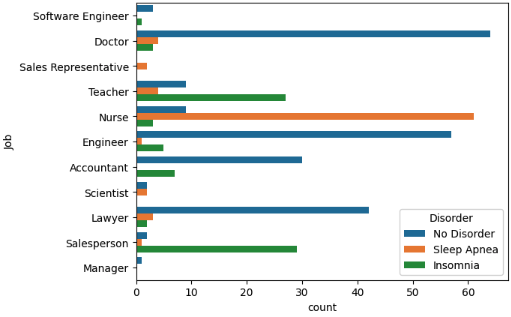
*Figure 1 – Gender x Sleep disorder*



The bar graph illustrates the distribution of sleep disorders among males and females, categorized into three groups: No Disorder, Sleep Apnea, and Insomnia. The 'count' on the y-axis represents the number of individuals within each category, providing a clear comparative view of sleep disorders between genders.

From the graph, it is evident that a significantly higher number of males report 'No Disorder' compared to females, with males nearly doubling the count of females in this category. For sleep disorders such as Sleep Apnea and Insomnia, males also show a higher prevalence compared to females in Sleep Apnea, but a slightly lower count in Insomnia. This suggests that while sleep apnea is more common among males, females are slightly more affected by insomnia. Overall, the graph highlights clear gender differences in the prevalence and type of sleep disorders, indicating potential biological, behavioral, or social factors influencing these patterns.

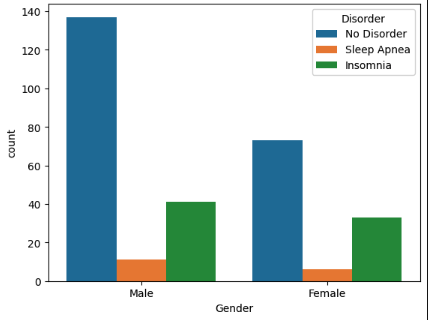
*Figure 2 – occupation x sleep disorder*



The bar graph presents the prevalence of different sleep disorders across various occupations, with disorders categorized into 'No Disorder,' 'Sleep Apnea,' and 'Insomnia.' Each profession is listed on the y-axis with corresponding counts of each condition illustrated through colored bars. This visualization allows for an easy comparison of sleep health among different job titles.

From the graph, it is noticeable that Nurses and Doctors have higher counts of 'No Disorder,' but Nurses also show significant instances of 'Insomnia.' In contrast, roles like Software Engineers, Sales Representatives, and Managers have a broader distribution of 'Sleep Apnea.' The graph also highlights that certain stressful or demanding professions, such as Lawyers and Salespersons, tend to show higher instances of 'Insomnia.' This pattern suggests that the nature of one's job, possibly including stress levels and work hours, could significantly influence sleep health and the type of sleep disorder experienced.

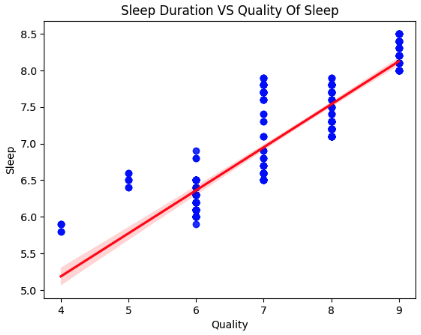
*Figure 3 – gender x sleep disorder excluding nurses*



The bar graph presents the prevalence of different sleep disorders across various occupations, with disorders categorized into 'No Disorder,' 'Sleep Apnea,' and 'Insomnia.' Each profession is listed on the y-axis with corresponding counts of each condition illustrated through colored bars. This visualization allows for an easy comparison of sleep health among different job titles.

From the graph, it is noticeable that Nurses and Doctors have higher counts of 'No Disorder,' but Nurses also show significant instances of 'Insomnia.' In contrast, roles like Software Engineers, Sales Representatives, and Managers have a broader distribution of 'Sleep Apnea.' The graph also highlights that certain stressful or demanding professions, such as Lawyers and Salespersons, tend to show higher instances of 'Insomnia.' This pattern suggests that the nature of one's job, possibly including stress levels and work hours, could significantly influence sleep health and the type of sleep disorder experienced.

*Figure 4 – sleep duration vs quality of sleep*



The scatter plot above illustrates the relationship between the quality of sleep and the duration of sleep among individuals. Each blue dot represents a specific data point that aligns with a person's sleep quality on the x-axis, ranging from 4 to 9, and their corresponding sleep duration in hours on the y-axis, varying from around 5.5 to 8.5 hours. The red line represents a trend line that indicates a clear positive correlation between the duration of sleep and its perceived quality. As the quality rating increases, so does the length of sleep, suggesting that better sleep quality is generally associated with longer sleep durations.

This visualization provides empirical support to the widely accepted notion that increased sleep duration is likely to enhance sleep quality. The density of points around the trend line, particularly for higher quality ratings (7 to 9), underscores a consistent pattern where individuals report greater satisfaction with their sleep as the duration extends. However, the presence of some scattered points also indicates variability, suggesting that while longer sleep generally correlates with higher quality, individual experiences can differ based on other factors such as health conditions, lifestyle, and stress levels.

*Figure 5 – stress level vs quality of sleep*



The scatter plot above depicts the relationship between stress levels and the quality of sleep, demonstrating a clear inverse correlation as depicted by the trend line. The x-axis represents the quality of sleep, ranging from 4 to 9, while the y-axis indicates the stress level on a scale from 3 to 9. The trend line slopes downward, indicating that as the quality of sleep improves, the reported stress levels decrease. This suggests that higher sleep quality is typically associated with lower levels of stress among individuals.

This visualization supports the hypothesis that stress management could be crucial in improving sleep quality. The spread of data points shows a general trend, but also highlights some variability, suggesting that while stress reduction is likely to improve sleep quality, the relationship may also be influenced by other factors such as individual health conditions, lifestyle choices, and environmental influences. As stress levels decrease, individuals generally experience better quality sleep, which could have wide-ranging positive effects on overall health and well-being.

Now testing the module for some of Machine Learning tests, I initially used the OLS Regression and got 0.791 accuracy. When I tested linear regression, I got the same value of 0.791. Those values are not a high accuracy, because it is under 80%.

The other test I did was kNN model, and according Prakash Shyam(2023), kNN model is a simple and effective machine learning algorithm for both classification and regression tasks. In KNN, the idea is to classify an unknown sample based on its distance to the K nearest samples in the training set.

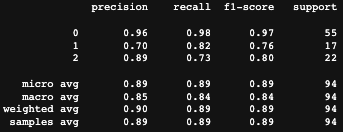
Using this test model, I got a table below and the results that I could extract from there were: in class 0 the precision was 96%, which is a good value. The recall was 98% and the F1-Score was 97%, which means the balance between Precision and Recall. The support for class 0 was 55, meaning the true instances for this class.

Classes with high precision ratio means there are many actually positive values and low false positive. In class 1 it is possible to see the precision was 70%, which is a lower value, not very accurate and the recall was 82% which is a bit better, but still not a satisfy result. The F1-Score for this class was 76% with 17 actual occurrences.

For class 2, the precision was a good result of 89%, but the recall was not accrue like the precision, getting a value of only 73%. F1-Score returns 80% and 22 actual occurrences, not a very satisfy value too.

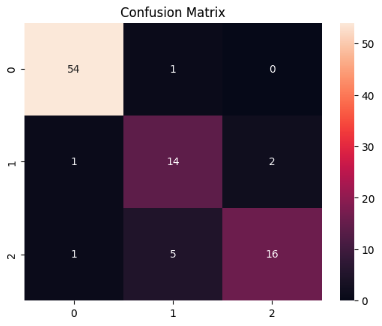
The other values, micro avg, macro avg, weighted avg and samples avg got values that can be considered ok, because it considered the three classes. Those values are around 89%, where is a good value, but if you look the class 0 are a bit lower, but if you compare with class 1 it is way higher.

*Figure 6 – table for kNN*



The confusion matrix have results similar with the kNN table, but with values bit lower than the previous test. The diagonal showing good values and the misclassified values are low, so it is satisfy result.

*Figure 7 – confusion matrix*



The confusion matrix is a visual tool used to evaluate the performance of a classification model by comparing the model's predicted results against the actual outcomes. The matrix is divided into rows that represent the actual classes and columns that represent the predicted classes, with each cell showing the count of predictions made by the model. In this specific matrix, there are three classes labelled as 0, 1, and 2. The diagonal cells (54 for class 0, 14 for class 1, and 16 for class 2) represent the number of correct predictions made by the model for each class, indicating true positives.

The off-diagonal cells show the number of incorrect predictions, where the model has misclassified the examples. For instance, class 0 has 1 instance misclassified as class 1, and class 1 has 2 instances misclassified as class 2 and 1 misclassified as class 0. Similarly, class 2 has 5 instances misclassified as class 1 and 1 as class 0. These misclassifications provide insights into the areas where the model may struggle, suggesting possible overlaps in the features of these classes or indicating areas where the model may require further training or adjustment to improve its accuracy.

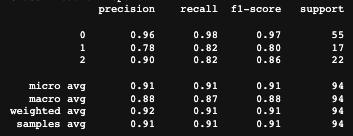
The effectiveness of the model can be assessed by examining the ratio of correct predictions (true positives) to the total predictions for each class, reflected in the diagonal of the confusion matrix. For instance, class 0 shows strong predictability with 54 correct predictions out of 55, indicating high reliability in identifying this class. Class 1 and class 2 have lower accuracy with 14 and 16 correct predictions out of 17 and 22, respectively, highlighting areas where the model's performance could be enhanced.

The presence of misclassifications between classes, particularly between classes 1 and 2, suggests that these classes share similar characteristics or that the features used by the model do not adequately distinguish between them. The analysis of such a confusion matrix is crucial in understanding the model's strengths and weaknesses, guiding further refinements and adjustments to improve classification accuracy, and ensuring the model is robust across various scenarios.

Testing the decision tree, which according SONG(2015), Decision tree methodology is a commonly used data mining method for establishing classification systems based on multiple covariates or for developing prediction algorithms for a target variable. This method classifies a population into branch-like segments that construct an inverted tree with a root node, internal nodes, and leaf nodes.

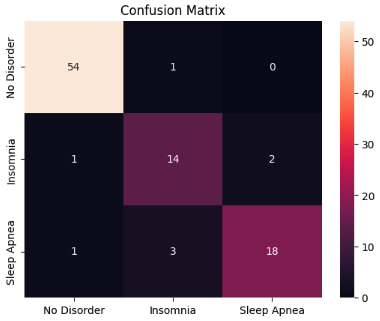
The classes 0 have the exactly same results of the kNN, but for class 1 and 2 they have increased a bit, which turns this more accrue than the previous test, as per table below.

*Figure 8 – decision tree table*



The other confusion matrix shows more accrue as well, showing a slide improvement if compared with the previous one. It seems a satisfactory result, with a small increase in accuracy compared to the previous test.

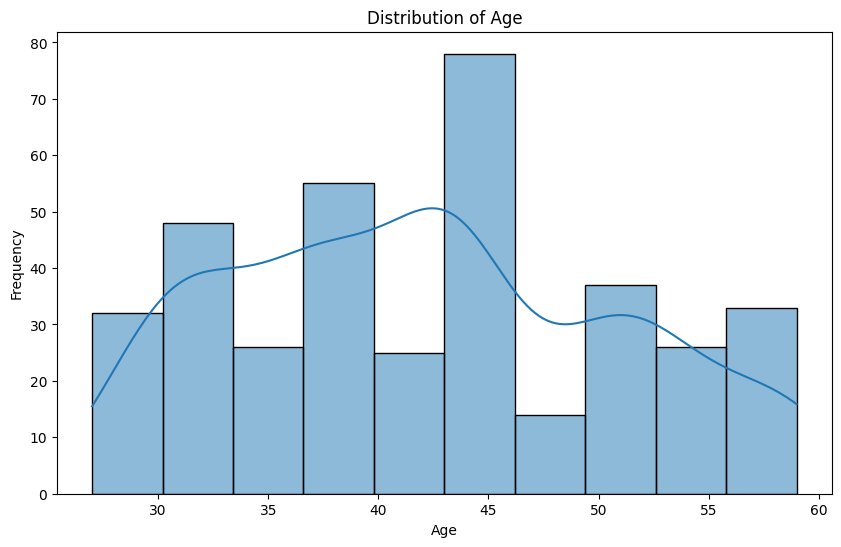
*Figure 9 – confusion matrix*



The graph displays the age distribution using a histogram and a KDE curve, offering insights into the frequency of various age groups. The age range 45-50 stands out as the most common, with around 80 individuals, followed closely by 35-40 and 40-45, each peaking at about 55-65 individuals. Conversely, the age groups 30-35, 50-55, and 55-60 have relatively low frequencies. The KDE curve reveals notable clustering around 35-40, 40-45, and 45-50, with a noticeable dip around 50-55.

Overall, the graph effectively portrays age distribution trends. The KDE curve provides a smooth overview of clustering patterns, aligning with the histogram's peaks and dips. For deeper insights, analyzing age alongside other demographic factors such as gender, education, or income could uncover underlying relationships. Additionally, more granular bin sizes could further illuminate subtle age distribution patterns.

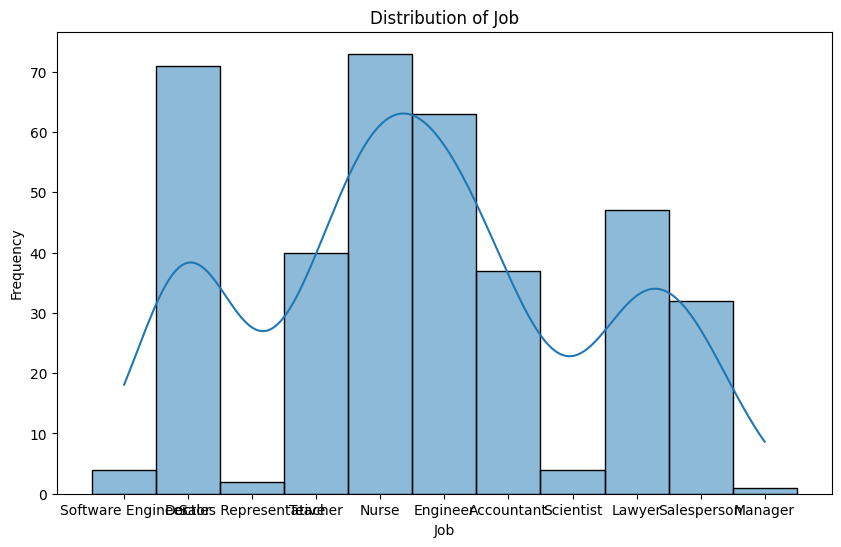
*Figure 10 – Distribution of age*



The graph illustrates the distribution of various jobs through a histogram, accompanied by a KDE curve to highlight the overall trends. The most common professions are Nurse and Sales Representative, each with frequencies around 70. They are followed closely by Engineer (65) and Lawyer (50), while other professions such as Teacher, Accountant, and Salesperson show moderate representation (around 30-40 individuals each). Professions like Software Engineer, Scientist, and Manager are less frequent, with counts below 10.

The KDE curve reveals two prominent peaks, indicating clustering around the professions Nurse and Sales Representative. The secondary peaks for Engineer and Lawyer further emphasize the popularity of these roles. However, the sharp dips between these peaks highlight underrepresentation in jobs like Software Engineer, Teacher, and Scientist. Overall, the graph effectively showcases trends in job distribution while also suggesting potential imbalances in certain professions. Further analysis could involve exploring correlations between job distribution and factors like education or geographic location for a more comprehensive understanding.

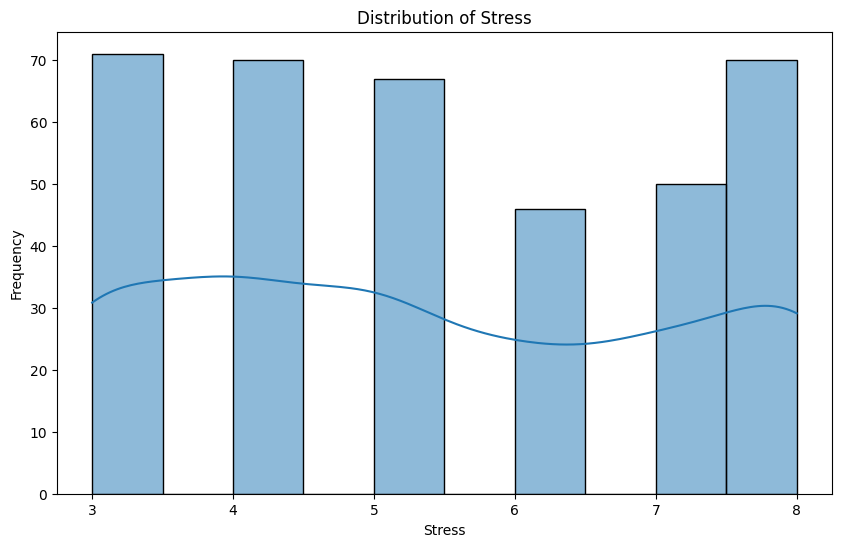
*Figure 11 – distribution of job*



The graph depicts the distribution of stress levels using a histogram, accompanied by a KDE curve. The stress levels are grouped into categories ranging from 3 to 8, and the frequency of individuals reporting each stress level is shown on the y-axis. Stress levels 3, 4, and 8 each show the highest frequencies, with around 70 individuals per category. Conversely, levels 5 and 6 are less common, with frequencies around 50 and 40, respectively. The KDE curve demonstrates a relatively stable pattern, with mild dips around stress levels 5 and 6 and a gradual rise toward levels 7 and 8.

Overall, the distribution suggests a prevalence of moderate to high stress levels among the individuals surveyed. The high frequencies for stress levels 3, 4, and 8 may indicate clustering around these specific points, perhaps due to the subjective nature of stress perception or external factors influencing stress categorization. The graph effectively highlights these trends and suggests a need to investigate potential causes of higher stress levels at certain points. Further analysis could include correlating stress levels with demographic data or exploring possible interventions to reduce stress.

*Figure 12- distribution of stress*

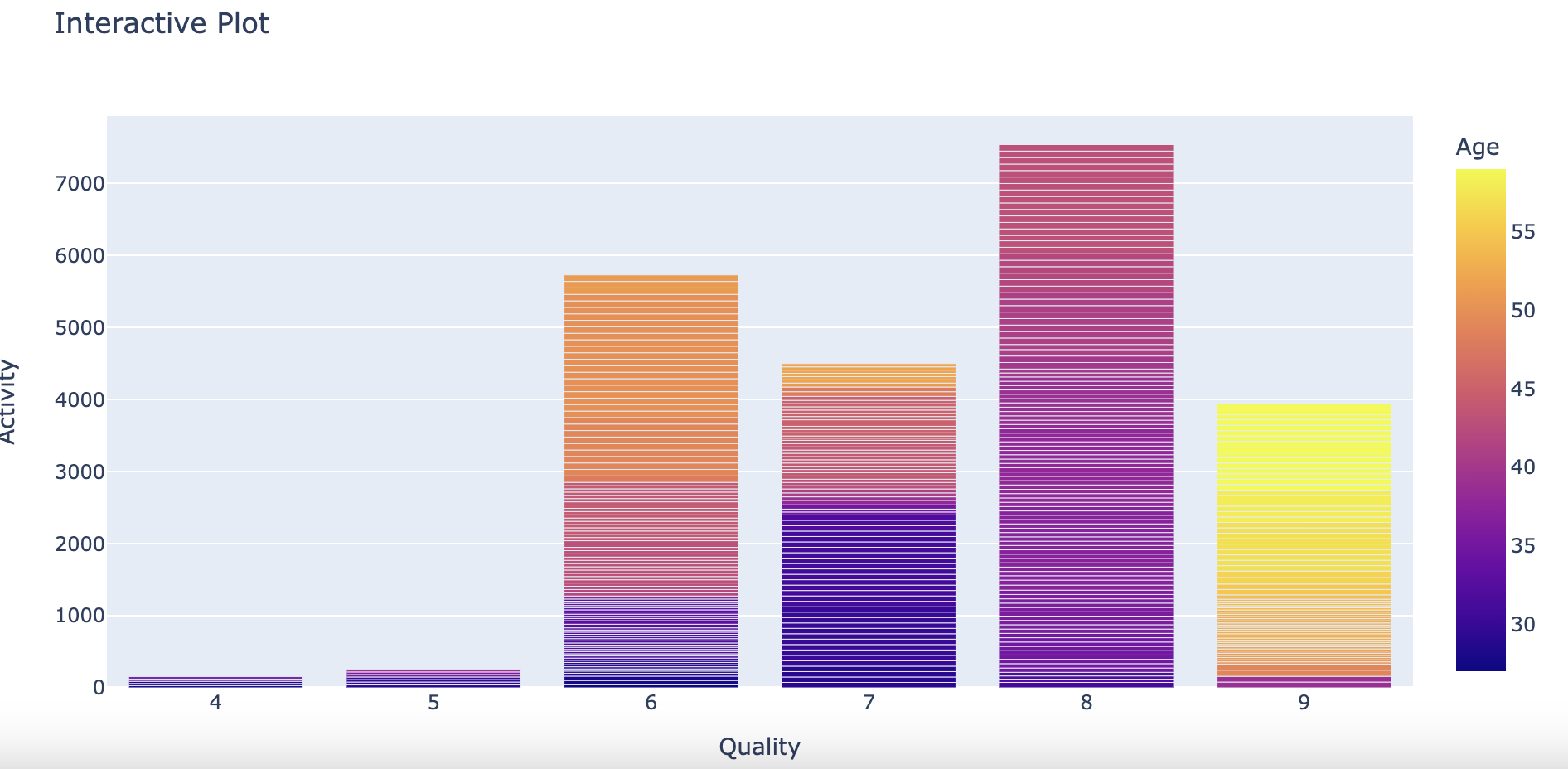


The graph provides an analysis of activity levels across different quality scores, incorporating age as a color-coded dimension. The x-axis represents quality of sleep scores ranging from 4 to 9, while the y-axis shows activity levels. The age gradient, ranging from 30 (purple) to 55 (yellow), is mapped on the right side of the graph. Quality scores 6 and 8 are the most frequently reported, with activity levels exceeding 7000 and 8000 units, respectively. Quality of sleep scores 7 and 9 follow behind with moderate activity levels around 4000-6000. In contrast, quality of sleep scores 4 and 5 have significantly lower activity, both below 500 units.

When analysing the colour gradient, it is evident that younger individuals (indicated by the darker purple shades) tend to be more active at higher quality of sleep scores like 8 and 9. On the other hand, individuals with moderate to lower activity levels and quality of sleep scores (6 and 7) appear to represent a more varied age range, with purple, yellow, and orange colours present. This suggests that middle-aged individuals are more distributed across varying activity levels.

Overall, the graph provides valuable insights into the relationship between age, activity levels, and quality of sleep scores. It indicates a clear trend where younger individuals are likely to achieve high quality scores with higher activity levels. Conversely, middle-aged and older individuals are more evenly distributed, contributing significantly to quality scores ranging from 6 to 8. The visualization effectively conveys the age distribution across different activity levels, making it useful for identifying demographic trends in relation to quality and activity.

*Figure 13 – quality of sleep x activity level*



The graph visualizes stress distribution across different professions, using a horizontal bar plot where each bar represents a specific job and is segmented by stress levels, ranging from 3 to 8 (indicated by the color scale on the right). The x-axis denotes the cumulative frequency of individuals, while the y-axis lists job titles. The color gradient from purple to yellow indicates increasing stress levels.

A noticeable trend is the high concentration of stress levels 3 and 4 (purple and yellow shades) among Doctors, Nurses, and Teachers, reflecting relatively low to moderate stress levels in these professions. However, a significant portion of individuals in these professions also report stress levels 5 and 6, indicating a distribution that covers a wide stress range. Notably, the Doctor profession shows the highest cumulative frequency, with over 500 individuals represented, followed by Sales Representative and Engineer.

On the other end of the spectrum, professions like Software Engineer, Manager, Scientist, and Lawyer show varying distributions, with stress levels 7 and 8 (red and orange shades) being more prominent. Specifically, Lawyers exhibit a higher occurrence of stress levels 7 and 8, pointing to potentially high-stress environments in this profession. The Salesperson role also shows a wide spread of stress levels, indicating a mixed experience among individuals in this field.

Overall, the graph provides a detailed overview of stress distribution across professions. It highlights the significant variance in stress levels even within the same job category, suggesting that individual factors may play a role in perceived stress. Furthermore, professions like Doctor and Nurse reveal high cumulative stress frequencies, emphasizing the need for stress management interventions in these high-pressure roles. The visualization effectively conveys the nuanced relationship between job type and stress levels.

*Figure 14 - job vs stress*



This capstone project explored the intricate relationship between sleep quality, stress levels, and various lifestyle factors using a synthetic dataset from Kaggle. By analyzing data on occupations, sleep disorders, sleep duration, stress levels, and physical activity, we aimed to identify patterns that could shed light on how these factors interact and affect each other. Our findings indicate that stress levels and sleep quality are inversely related, with certain professions showing a clear link to higher stress levels and poorer sleep quality. Occupations that demand longer working hours or have lower wages, like nursing and sales representation, tend to exhibit elevated stress levels and an increased prevalence of sleep disorders.

The machine learning models implemented in this project, including OLS Regression, Linear Regression, k-Nearest Neighbors (kNN), and Decision Trees, provided insights into the predictability of sleep quality and stress levels. While the initial accuracy scores for the OLS and Linear Regression models were satisfactory but below the desired threshold, the kNN and Decision Tree models showed promise in their ability to classify and predict stress levels based on a range of features. The Decision Tree model, in particular, demonstrated improved classification accuracy compared to kNN, providing a robust framework for understanding stress-related factors.

Visualization techniques, such as histograms, KDE curves, heatmaps, and bar plots, were instrumental in uncovering trends in the data. For instance, the age distribution revealed clustering around certain age groups, particularly 35-50, while the job distribution highlighted the prominence of specific professions like Nurse and Sales Representative. Furthermore, the correlation analysis showed that stress levels are higher among individuals with shorter sleep durations and poorer sleep quality. The insights derived from this analysis can help inform interventions to promote better sleep hygiene and stress management.

In summary, this project underscored the significant impact of lifestyle factors on sleep quality and stress levels. It highlighted the importance of addressing occupational stressors, encouraging healthier habits, and identifying individuals at higher risk for sleep disorders. While the models and analyses provide a valuable foundation, further research could explore more complex interactions between factors and validate these findings with real-world data. Ultimately, the predictive model aims to empower individuals and organizations with actionable insights to improve well-being and reduce stress-related health issues.

The link for my GitHub repository is [here](https://github.com/ArthurVerza/StrategicThinking.CA2)

**Reference list**

**Jansen, E. (2020). *Sleep 101: Why Sleep Is So Important to Your Health | The Pursuit | University of Michigan School of Public Health | Adolescent Health | Child Health | Chronic Disease | Epidemic | Mental Health | Obesity*. [online] sph.umich.edu. Available at: https://sph.umich.edu/pursuit/2020posts/why-sleep-is-so-important-to-your-health.html.**

**Müller, A.C. and Guido, S. (2016). *Introduction to machine learning with Python : a guide for data scientists*. First Edition ed. Beijing: O’reilly, p.25.**

Ramar, K., Malhotra, R.K., Carden, K.A., Martin, J.L., Abbasi-Feinberg, F., Aurora, R.N., Kapur, V.K., Olson, E.J., Rosen, C.L., Rowley, J.A., Shelgikar, A.V. and Trotti, L.M. (2021). Sleep is essential to health: an American Academy of Sleep Medicine position statement. Journal of Clinical Sleep Medicine, [online] 17(10). doi:https://doi.org/10.5664/jcsm.9476.

Prakash Shyam, Karuppiah . “K Nearest Neighbor - an Overview | ScienceDirect Topics.” www.sciencedirect.com, 2023, [www.sciencedirect.com/topics/biochemistry-genetics-and-molecular-biology/k-nearest-neighbor](http://www.sciencedirect.com/topics/biochemistry-genetics-and-molecular-biology/k-nearest-neighbor).

Song, Yan-Yan, and Ying Lu. “Decision Tree Methods: Applications for Classification and Prediction.” Shanghai Archives of Psychiatry, vol. 27, no. 2, 2015, pp. 130–5, https://doi.org/10.11919/j.issn.1002-0829.215044.