Sleep, Lifestyle, and Health of Professionals and Students

By: Team Sleeping On It

Team Members:
Aarav Gadkar (aaravg@stanford.edu)
Ishaan Adhikary (ishaan07@stanford.edu)
Justin Ma (jma2026@stanford.edu)
Mohammad Umar (umar499@stanford.edu)

Discussion Section: **DIS-03**

Introduction

The purpose of this project is to determine the trends of sleep among general populations and students at Stanford, especially in determining the factors that have the biggest impact on sleep length and quality. With this information, suggestions to improve sleep habits and diagnose sleep disorders could be formed. To these ends, three datasets were collected: a Kaggle dataset regarding sleep, limited health information, and lifestyle choices for various professional adults; a second Kaggle dataset regarding sleep in more detail along with much more in-depth health information across a diverse group; and a final dataset constructed by our group detailing some lifestyle decisions and sleep data from students at Stanford. All of these datasets cover different data about different groups.

Questions throughout this document mainly answered questions about predicting sleep quality, disorders, and scores, but other questions such as determining the difference between different demographics, grouping students and drawing conclusions from the results, and building interactive models to offer patients who are struggling with getting healthy sleep direct and personalized suggestions. People who are not able to get healthy sleep usually end up lethargic and with weak immune systems, causing danger to themselves, but also sometimes to others through car accidents and other sleep-deprived and potentially lethal mistakes. By collecting data and improving suggestions for patients unable to get adequate sleep, patients will generally become happier and healthier.

Data Description

Data Set 1: Sleep, Health, and Lifestyle:

https://www.kaggle.com/datasets/henryshan/sleep-health-and-lifestyle

This dataset, containing 373 rows and 13 columns, highlights various variables associated with sleep and lifestyle data. Each entry in this dataset represents one individual, represented as an ID, containing **basic information** such as their age, gender, and occupation. It also includes an overview of the individual's **lifestyle**, including physical activity, stress level, and daily steps, as well as common **vital signs**: BMI category, blood pressure, and heart rate. Finally, this dataset includes the individual's **sleep habits** by measuring sleep duration, quality of sleep, and the presence of a sleep disorder.

Data Description (Dictionary pulled from Kaggle website)**:**

General Information:

- Person ID (Other): A unique identifier for each participant.
- Gender (Categorical): The participant's gender (Male/Female).
- Age (Numeric): The participant's age in years.
- Occupation (Categorical: The job or career of the participant.

Lifestyle:

- Physical Activity Level (minutes/day) (Numeric): The daily duration of physical activity for the participant, measured in minutes.
- Stress Level (scale: 1-10) (Numeric): A subjective assessment of the participant's stress level on a scale from 1 to 10.
- Daily Steps (Numeric): The number of steps the participant takes each day.

Vital Signs:

- BMI Category (Categorical): The participant's BMI classification (e.g., Underweight, Normal, Overweight).
- Blood Pressure (systolic/diastolic) (Numeric): The participant's blood pressure, represented as systolic pressure over diastolic pressure.
- Heart Rate (bpm) (Numeric): The participant's resting heart rate, measured in beats per minute.

Sleep Measurements:

- Sleep Duration (hours) (Numeric): The daily sleep duration of the participant in hours.

- Quality of Sleep (scale: 1-10) (Numeric): A subjective assessment of sleep quality on a scale from 1 to 10.

Sleep Disorder (Categorical): The presence or absence of a sleep disorder in the participant (None, Insomnia, Sleep Apnea).

Data Modifications:

- Set index to “Person ID”
- Fill “NaN” values with “None” in “Sleep Disorder” column
- Combine redundant occupations (e.g. Sales Representative and Salesperson)
- Combine redundant BMI categories (Normal and Normal Weight)
- Split blood pressure data into separate systolic and diastolic data.

Data Set 2: Sleep Efficiency:

https://www.kaggle.com/datasets/equilibriumm/sleep-efficiency
This dataset, containing 452 rows and 15 columns, highlights various variables associated with sleep patterns and lifestyle choices data. Each entry in this dataset represents one individual, represented as an ID, containing **basic information** such as their age and gender. It also includes an overview of the individual's **lifestyle choices**, including caffeine consumption, alcohol consumption, smoking status, and exercise frequency. This dataset measures **sleep patterns** by measuring sleep duration (which includes the specific time of day for bedtime and wake up time), sleep efficiency, the percentages of time spent in different stages of sleep (light sleep, deep sleep, and REM sleep), and finally number of awakenings during the night.

For this dataset, we are proposing to calculate the quality of sleep based on all the different sleep patterns (and also age). We found this article from Healthline to help us understand how to do this: https://www.healthline.com/health/how-much-deep-sleep-do-you-need. The specific formula is described in question 1.

Data Description (Dictionary pulled from Kaggle website)**:**
Basic Information:

- ID (Other): A unique identifier for each test subject.
- Gender (Categorical): Male or Female.
- Age (Numeric): Age of the test subject.

Lifestyle Choices:

- Caffeine Consumption (Numeric): The amount of caffeine consumed in the 24 hours prior to bedtime (in mg)
- Alcohol Consumption (Numeric): The amount of alcohol consumed in the 24 hours prior to bedtime (in oz)
- Smoking Status (Categorical): Whether or not the test subject smokes
- Exercise Frequency (Numeric): The number of times the test subject exercises each week

Sleep Patterns:

- Bedtime & Wake Up Times (Other): Includes dates and timestamps of when the test subject goes to bed and wakes up
- Sleep Duration (Numeric): The total amount of time the test subject slept (in hours)

- Sleep Efficiency (Numeric): A measure of the proportion of time in bed spent asleep
- REM, Deep, and Light Sleep Percentages (Numeric): The percentage of total sleep time spent in each stage of sleep
- Awakenings (Numeric): The number of times the test subject wakes up during the night

Data Modifications:

- Set index to “Person ID”
- Remove date information from bedtime and wakeup times

Data Set 3: Stanford Summer Students Sleep Survey:

https://github.com/gaminghusky1/DataScienceProject/blob/main/StanfordSleepSurvey.csv

This dataset, containing 42 rows and 11 columns, highlights various variables associated with sleep patterns, academic factors, and lifestyle. The dataset was constructed by our group through a questionnaire where students from the Data Science class and Trancos house during the 2024 Summer Session were encouraged to fill out responses. Each entry in this dataset represents one student containing **demographic information** such as their age and gender. It also contains **sleep information**, mainly through a subjective rating provided to us about their sleep. In addition, each response contains **academic information** such as units taken at Stanford and steps taken in a day. Finally, **health information** like steps and sleep or mental health disorders were provided by respondents.

Data Description

Metadata:

- Timestamp (Other): The date and time the survey response was received. This is unneeded for our analysis.

Demographic Information:

- Gender (Categorical): Male or Female.
- Age (Numeric): Age of the student, in years.

Sleep Information:

- Quality of Sleep (Numeric): A subjective assessment of sleep quality on a scale from 1 to 10.

Academic Information:

- Class Units (Numeric): The amount of units the student took during the Summer Session at Stanford. According to Stanford, each unit represents approximately 3 hours of work per week.

Lifestyle Information:

- Method of Transport (Categorical): One of Bike, Walk, or Drive. How this student most often travels to their classes.
- Stress Level (Numeric): A subjective assessment of stress on a scale from 1 to 10.
- Steps (Numeric): The amount of steps this student takes on average per day.
- Notes (Other): Miscellaneous remarks from each respondent. These will be unneeded in our analysis.

Health Information:

- Sleep Disorder (Text): A description of medical sleep disorders such as Insomnia provided by the student. Keywords can be used to categorize each response.
- Mental Health Disorder (Text): A description of medical mental health disorders such as Anxiety or Depression provided by the student. Keywords can be used to categorize each response.

Data Modifications:

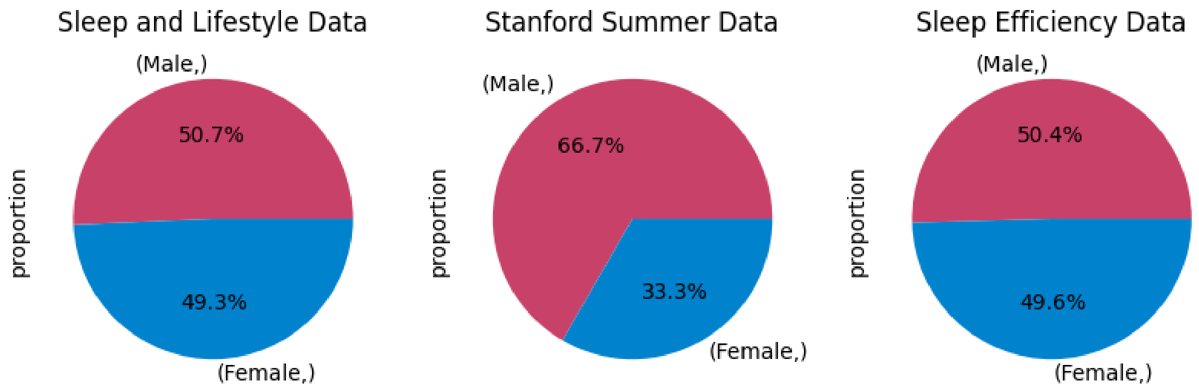
- Format responses in the “Steps” column to be numerical by stripping unnecessary characters like commas and interpreting ranges marked by dashes.
- Vectorize keywords from the Health Information columns “Sleep Disorder” and “Mental Health Disorder” such as Insomnia and Sleep Apnea or Anxiety, Depression, Obsessive-Compulsive Disorder (OCD), and Depersonalization-Derealization Disorder (DPDR).
- Drop columns that are not meaningful, like the timestamp of when the survey response was recorded.

Exploratory Data Analysis

Generally, each dataset focuses on wildly different age groups:

- Dataset 1 has a mean age of 42 and standard deviation of about 8.6, with a minimum age 27 and maximum age of 59. This indicates that the dataset generally has an even spread of data concerning adults in the middle age range, beginning at those leaving early adulthood and ending at those approaching seniority.
- Dataset 2 has a mean age of 40 and a standard deviation of about 13.2, with a minimum age of 9 and a maximum age of 69. This indicates that the dataset generally has an even spread of data on people of many different age groups, including children and teens, along with older people who may suffer more often from sleep conditions.
- Dataset 3 is the black horse of the group, focusing mostly on teenagers and young adults, with a mean age of 17 and a standard deviation of about 2.5, with a minimum age of 16 and a maximum age of 29. This dataset is much more focused in regards to age, and mostly concerns older teens and young adults, with minimal crossover in age ranges with the other two datasets.

Further, analyzing the makeup of biological sex across datasets is helpful in identifying possible biases in the collected data:



Sleep and Lifestyle Data

Stanford Summer Data

Sleep Efficiency Data

Overall, the datasets are generally very well balanced, although Dataset 3 (pictured in the middle) generally overrepresents males and underrepresents females.

Dataset 1

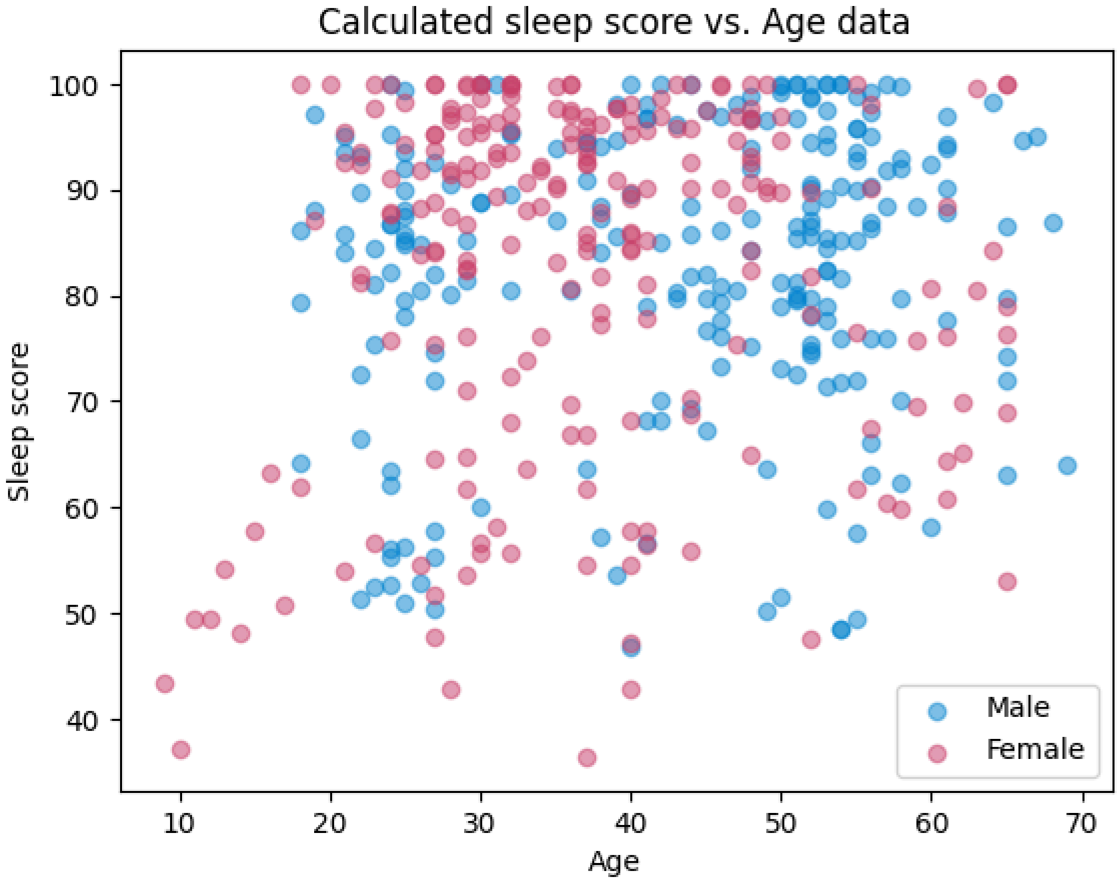
All EDA for Dataset 1 was shown within the questions for ease of reference.

Dataset 2

No EDA was used for question 1

EDA for question 2

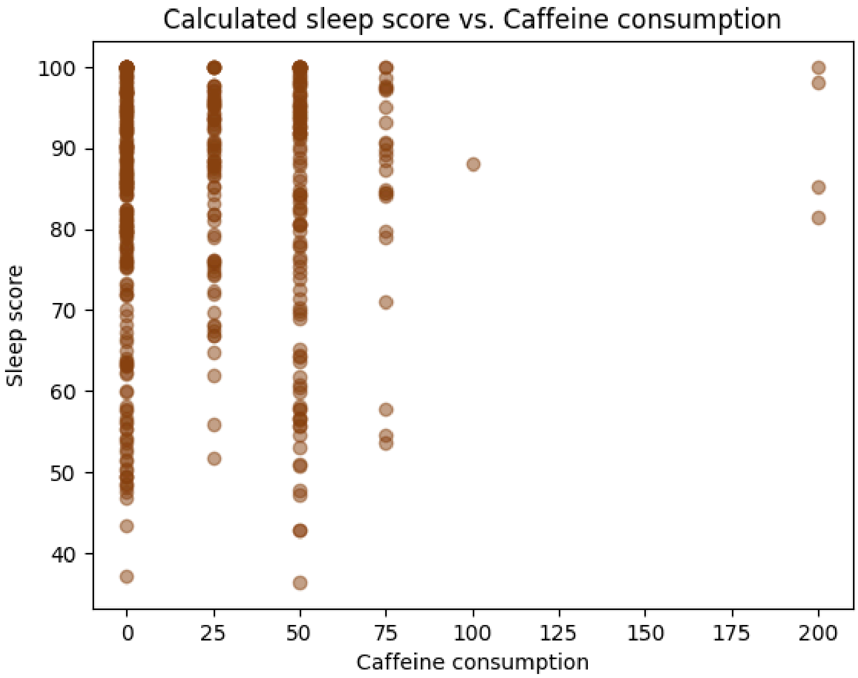
Below is a graph that shows the correlation between an individual's age, biological sex, and their calculated sleep score (using the formula in question 1).

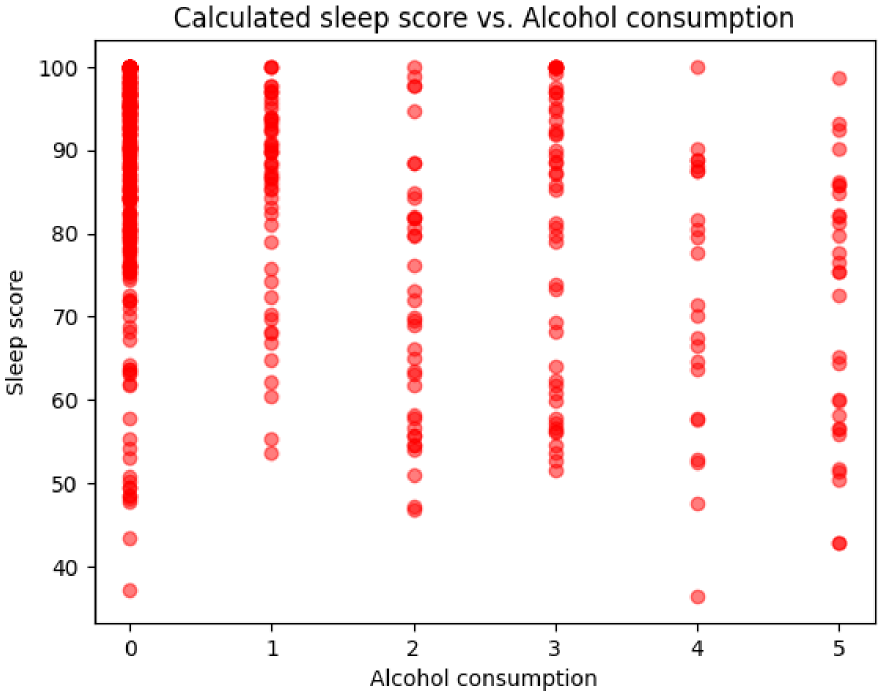


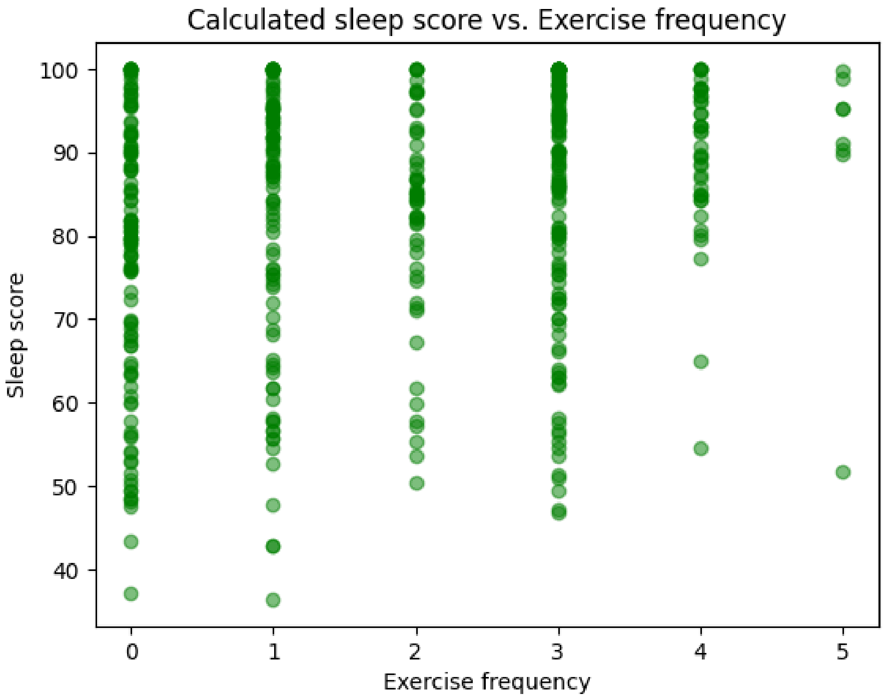
Based on the spread of the data points, there does not appear to be a significant correlation between age and sleep score, or between biological sex and sleep score. Based off of this graph, we can see that these features will probably not be good for predicting sleep score.

EDA for question 3

Below are three graphs that show the correlations between lifestyle habits and sleep score. The three lifestyle habits are: Caffeine consumption in the 24 hours before bedtime (mg), Alcohol consumption in the 24 hours before bedtime (oz), and number of days of exercise per week.


Calculated sleep score vs. Caffeine consumption
Sleep score
100
90
80
70
60
50
40
0
25
50
75
100
125
150
175
200
Caffeine consumption

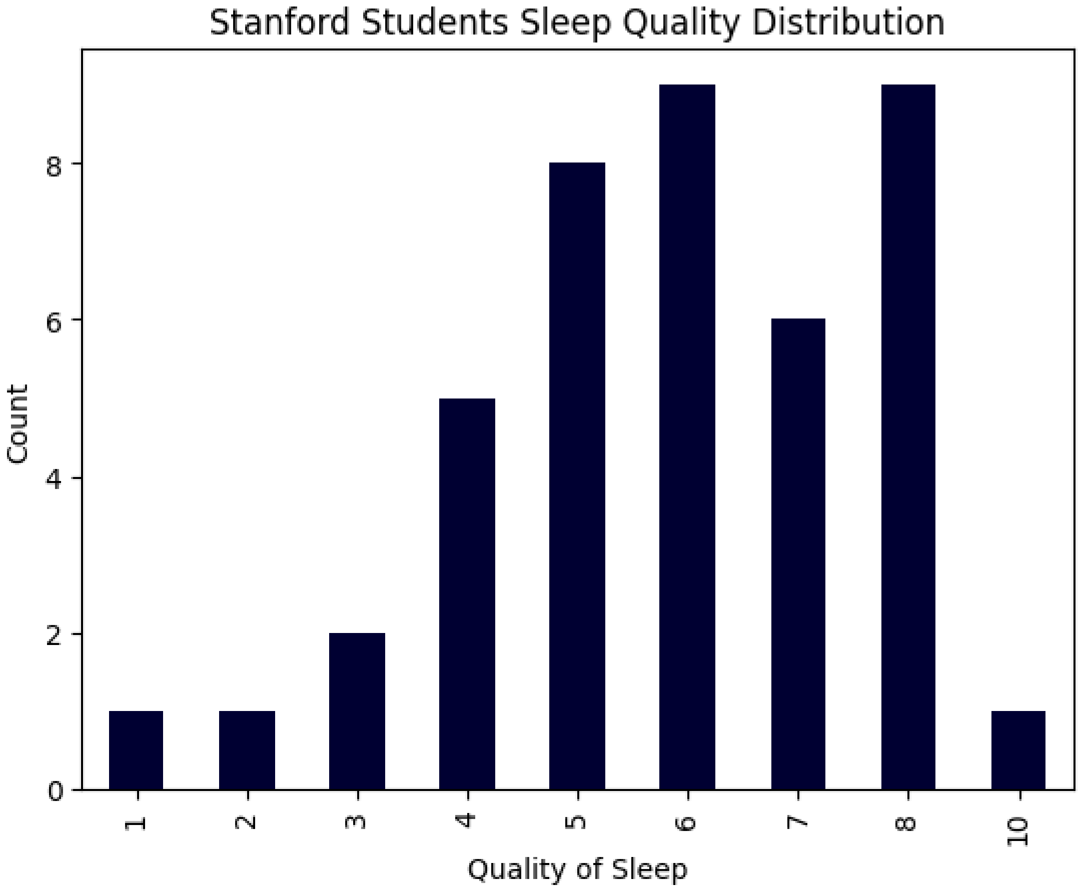

Calculated sleep score vs. Alcohol consumption
Sleep score
100
90
80
70
60
50
40
0
1
2
3
4
5
Alcohol consumption



We expected there to be a fairly strong negative correlation between caffeine consumption and the sleep score, but there seems to be no negative correlation at all based on the graph. In fact, the correlation between alcohol consumption and sleep score seems stronger than the correlation between caffeine consumption and sleep scores based on its graph. There also does not seem to be a very strong correlation between exercise frequency and sleep score, and they have a weak positive correlation.

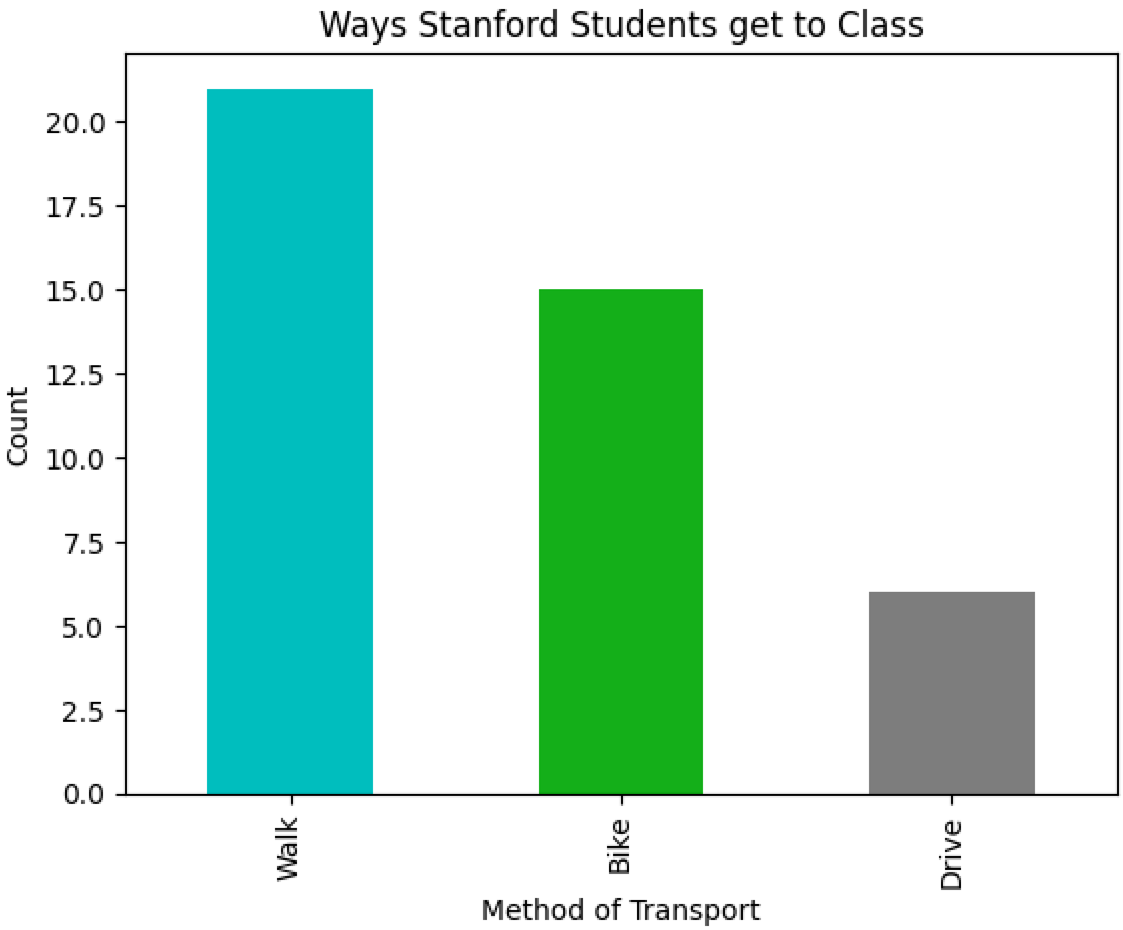
Dataset 3

One useful piece of information to know is how well-representative our dataset is of students with varying qualities of sleep. One good way to observe this is to plot the responses received for sleep quality in a bar graph.



Stanford students recorded in our data tend to get decent but not amazing and not terrible sleep, but we have samples for every rating, including terrible sleepers and amazing sleepers, except for students who rate their sleep as a 9. The vast majority of students rated their sleep as either a 6 or an 8. Overall, models trained on this data may have trouble predicting lower or higher values for the quality of sleep, as they are generally underrepresented in this data set. It is likely models trained on this data will devolve into a mean predictor.

Finally, understanding the spread of how students get to class will likely also be a useful metric.



Overall, students are generally active and walk or bike to class rather than driving. This is likely due to the generally low age of students at the Summer Session, and the fact that for students who do not live at Stanford bikes and cars cost money to rent. This fact is useful in interpreting the steps variable.

Project Design, Implementation, and Results

Data Set 1 Question 1

Analytical Question:
Understanding the data: what is the relationship between sleep duration, quality of sleep, and sleep disorders?

Comments:
In this dataset, we treat sleep duration, quality of sleep, and sleep disorders as “outputs” that we will ultimately center the models we build in the future questions around. We want to find the relationships, if any, between the three variables. This will help us better understand the data, and will help develop models in future questions.

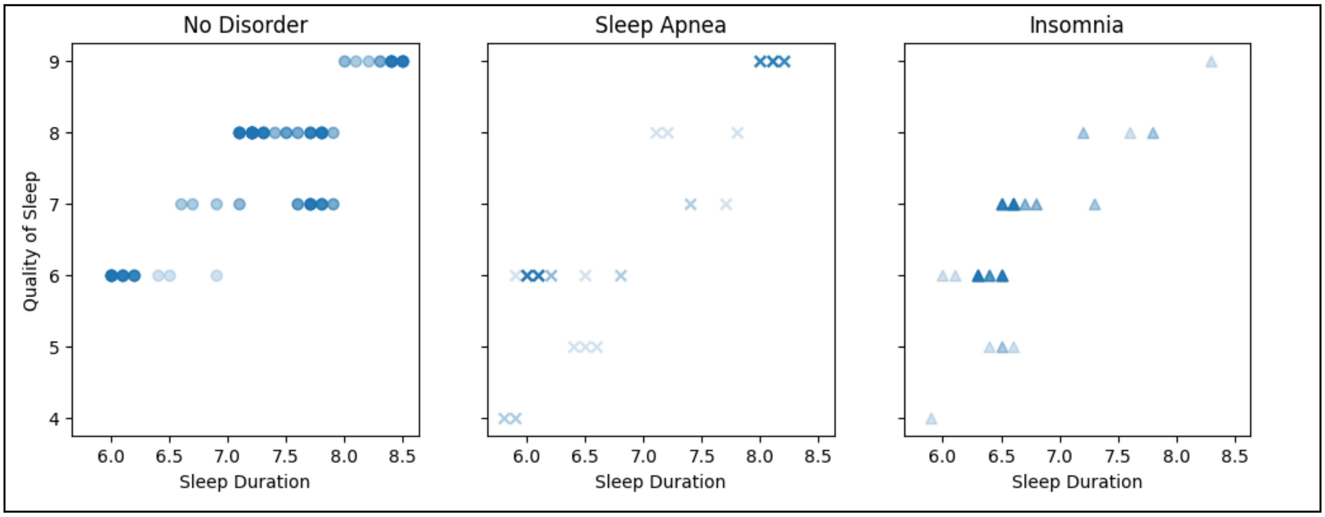
Data Used:
For this question, we will be looking at the “Sleep, Health, and Lifestyle” dataset, specifically the sleep measurements (sleep duration, quality of sleep, and sleep disorders).

Proposed Solution:
The solution is that we make various types of plots, plotting each of these variables against each other to find relationships. We can also use cross tabulation and value counts functionality to make some data frames that further help us understand these variables. We can use clustering techniques to understand groups of data.

Proposed evaluation:
Ultimately, what we get from this data analysis are multiple figures that help us understand the data. As stated in the comments, this will help us understand what to use to build our models. Though we will most likely still use all three variables throughout the data analysis, it will give us a clear understanding of our results.

Results:

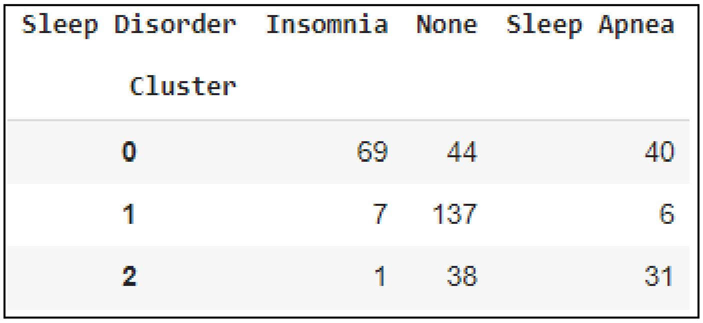
Scatter plot of sleep duration vs. sleep quality for varying sleep disorder types.



We can see that sleep duration and quality of sleep are pretty strongly correlated, with a correlation coefficient of ~0.88.

From the plots, we can see that if you don't have a sleep disorder, your quality of sleep is 6 or above, but if you do have a sleep disorder, you can have a quality of sleep as low as 4. This makes sense, because people with sleep disorders often have trouble sleeping or frequent interruptions in the nighttime, which may decrease quality of sleep.

From the clustering model we created, we can see that there are no clear clusters we can evaluate from sleep duration and quality of sleep. Looking at the distributions of sleep disorders over the modeled clusters, we can see that they are pretty spread out. This makes sense because, looking at the plots, even people with sleep disorders may have good quality of sleep and high sleep duration (likely due to medical treatments to improve these disorders). We will evaluate other factors in future questions in order to find which impacts sleep habits.



Sleep Disorder	Insomnia	None	Sleep Apnea
Cluster			
0	69	44	40
1	7	137	6
2	1	38	31

In conclusion, for the following questions, we will use quality of sleep as the feature to evaluate our predictive models. We will also consider sleep disorder as the categorical feature.

Data Set 1 Question 2

Analytical Question:
What are the effects of a person's age, gender, and occupation on their sleep duration and quality of sleep?

Do these factors (age, gender, and occupation) have a strong influence on sleep disorders?

Comments:
This question helps us understand whether or not an individual's characteristics (which for the purposes of this question includes occupation) has a significant influence on sleep habits. We can find different statistics between these two sets of variables such as probabilities and relationships using pandas functionalities.

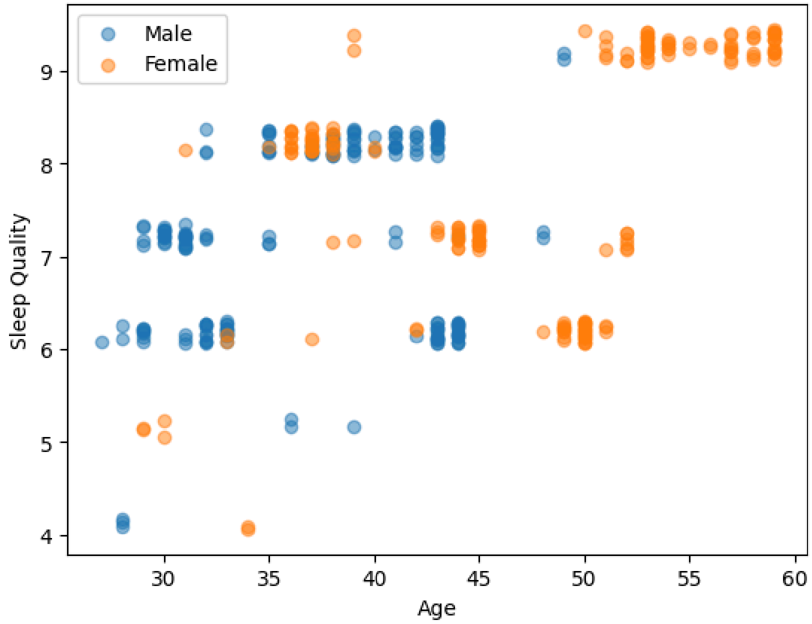
Data Used:
The data we are going to use for this question is the "Sleep, Health, and Lifestyle" dataset. We are going to look at the person's characteristics (including occupation) as well as the sleeping habits and sleep disorders.

Proposed Solution:
We can use pandas functionalities like cross tabulation, grouping, and plotting to find different relationships. This can help us answer certain questions relying on things like joint distributions, conditional probabilities, and different proportions. For example, we could ask questions such as, "Out of all male engineers, how many have Sleep Apnea but good sleep habits." Understanding questions like these will help us get a good grip on the data we are analyzing. Using a linear regression model to predict sleep habits to see how the features affect them.

Proposed Evaluation:
From this analysis, we can see trends and relationships in our dataset that relate characteristics to the features of sleep.

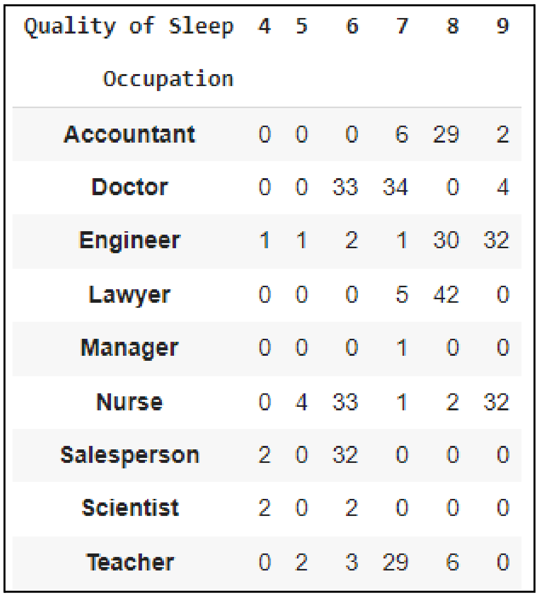
Results:
Because of our original analysis, we will be simplifying this question to only look at quality of sleep because of the correlation between quality of sleep and sleep duration. We did both exploratory data analysis, then used parametric models in order to solidify our findings.

Jittered scatter plot of age vs. sleep quality for males and females



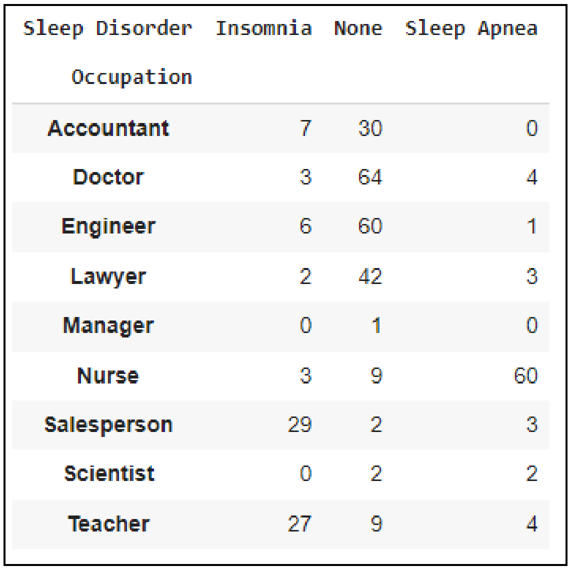
From the plot, we can see that there are a large number of older women that have very good quality of sleep and a large chunk of people between 35-45 that have good sleep quality. What is surprising is that younger people tend to have worse sleep quality than older people.

Distributions of quality of sleep across occupations



Quality of Sleep	4	5	6	7	8	9
Occupation						
Accountant	0	0	0	6	29	2
Doctor	0	0	33	34	0	4
Engineer	1	1	2	1	30	32
Lawyer	0	0	0	5	42	0
Manager	0	0	0	1	0	0
Nurse	0	4	33	1	2	32
Salesperson	2	0	32	0	0	0
Scientist	2	0	2	0	0	0
Teacher	0	2	3	29	6	0

Distributions of sleep disorders across occupations



Sleep Disorder	Insomnia	None	Sleep Apnea
Occupation			
Accountant	7	30	0
Doctor	3	64	4
Engineer	6	60	1
Lawyer	2	42	3
Manager	0	1	0
Nurse	3	9	60
Salesperson	29	2	3
Scientist	0	2	2
Teacher	27	9	4

From our analysis of occupation, we can see that lawyers and engineers tend to have better sleep quality than others, and salespersons, teachers, and doctors tend to have mediocre sleep quality. We can also see that nurses tend to have sleep apnea, and salespersons and teachers tend to have insomnia.

Linear Regression Model Data

Feature	**Coefficient**
Age	0.752
Female	0.048
Male	-0.048
Accountant	0.768
Doctor	0.220
Engineer	0.696
Lawyer	0.877
Nurse	-0.830
Salesperson	-1.408
Teacher	-0.322

Looking at the coefficients of the linear regression model, the coefficient for age is a larger value, which shows that it has a large impact on quality of sleep. Gender does not have much impact on quality of sleep because their coefficients are very small. The impact on quality of sleep for each occupation is shown through each individual coefficient. For example, the salesperson coefficient, -1.408, shows that salespersons generally have bad quality of sleep.

Feature	**Coefficient (Insomnia)**	**Coefficient (None)**	**Coefficient (Sleep Apnea)**
Age	0.553	-0.711	0.158
Female	-0.373	0.396	-0.022
Male	0.374	-0.395	0.022
Accountant	0.674	0.281	-0.955
Doctor	-0.704	0.704	0
Engineer	-0.458	1.365	-0.907
Lawyer	-0.936	1.029	-0.093
Nurse	-1.216	-0.912	2.129
Salesperson	1.252	-1.226	-0.026
Teacher	1.388	-1.240	-0.148

To evaluate these results, we will be saying that any intercept with a magnitude of 1 or greater shows a strong impact on sleep disorder.
The coefficients of insomnia shows that salespersons and teachers are often affected because they have positive coefficients greater than 1. It also shows a negative coefficient with a magnitude greater than 1, which shows that nurses will rarely suffer from insomnia. Similarly, we see in the coefficients for no sleep disorder that engineers and lawyers are generally predicted to have no sleep disorder, while salespersons and teachers often do have a disorder. In the coefficients for sleep apnea, we see a large coefficient for nurses, showing that they often have sleep apnea. This information is consistent with our cross tabulation above.

Data Set 1 Question 3

Analytical Question:
How does lifestyle (stress level, physical activity, and daily steps) and vital signs (BMI, blood pressure, and heart rate) influence quality of sleep?

Can we provide suggestions to change one's lifestyle habits to improve sleeping habits?

Are these suggestions different for males and females?

Comments:
In the last question, we observed how features that are more or less out of our control influence sleep habits. In this question, we want to see how things we can change, being lifestyle and vital signs (this can be through getting treatments or taking better care of yourself), influence quality of sleep. This can be used to give suggestions to improve an individual's sleeping habits by evaluating what critical factors in the lifestyle can be changed. We can also do this separately for males and females just to see if there are any significant changes in our results to get a clearer understanding of who needs to change what to improve their sleep.

Data Used:
For this question, we are using the "Sleep, Health, and Lifestyle" dataset, specifically the features mentioned above, including gender and both systolic and diastolic parts of blood pressure. We are only looking at quality of sleep for this question because we want to analyze these factors against something out of the individual's control (we are not looking at sleep duration because you can go to sleep whenever you would like regardless of your health and lifestyle). Use a standard scaler to scale the data.

Proposed Solution:
For this question, we can use linear regression to examine the critical features within lifestyle and vital signs that impact quality of sleep. We can also use KNN. To get a model that we have confidence in, we will use cross validation for both KNN and linear regression.

Proposed Evaluation:
The coefficients for the resulting linear model will give us insights to which features are most important to improve quality of sleep. Using this, we can give individuals specific suggestions to improve their quality of sleep by changing their lifestyle. The results may

be different depending on your gender, so for a more in-depth analysis, we will do this process independently for each gender.

Results:

RMSE Table

Male		**Female**	
Linear Regression	**KNN**	**Linear Regression**	**KNN**
0.362	0.326	0.685	0.230

Overall, both prediction models worked well to predict quality of sleep based on lifestyle and vital signs. The KNN model was more accurate, but linear regression was not far off either. However, the KNN model does not offer any quantitative way to evaluate the impacts of the different features on quality of sleep, we can trust the coefficients of the linear regression models.

Feature Coefficients (Linear Regression)

Male		**Female**	
Stress Level	-0.738	Stress Level	-1.120
Physical Activity	0.062	Physical Activity	0.285
Heart Rate	0.007	Heart Rate	-0.081
Diastolic	0.154	Diastolic	0.280
Normal Weight	0.735	Normal Weight	0.565
Overweight	-0.702	Overweight	-0.231
Obese	-0.033	Obese	-0.334

In general, we can see that stress level has a negative impact on sleep quality, and also that it is important to try to stay at a normal weight range, as it positively affects your sleep quality, while being overweight or obese negatively affects your sleep quality.
For males specifically, it seems that weight is more impactful on their sleep quality than it is for females. And for females, it seems that stress level is more impactful on their sleep quality than it is for males. This information would come in handy in the real world when trying to help a certain patient overcome problems in their sleep habits.

Data Set 1 Question 4:

Analytical Question:
Predict the **probability** of having a sleep disorder based on the following features:

- **Characteristics:** Age, Gender
- **Lifestyle:** Stress level, Physical activity, Daily steps
- **Vital Signs:** BMI, Blood pressure, Heart rate

Which of the features have a stronger influence on the probability of having a sleep disorder?

Comments: In this question, we are trying to predict the probability of having a sleep disorder based on the above mentioned features. To do this, we will evaluate the attributes of a logistic regression model that we create trained on these features. However, we will eliminate some of these features that are highly correlated, and we will also convert all the categorical variables into numeric variables using binary 1s and 0s. For simplicity sake, we will be considering sleep apnea and insomnia as equivalent values of sleep disorders.

Data Used: We will be using the "Sleep, Health, and Lifestyle" dataset for this question, and implementing all features mentioned in the question, including sleep disorder as an output variable. We will be leaving out sleep duration and quality of sleep because it is not reasonable to include these for newer patients who may not have measured them. We are excluding occupation because we have already analyzed the influence of occupation in a previous question, so we can simplify the analysis done in this question.

Proposed Solution: We will use logistic regression with the simplified features and evaluate the coefficients and intercepts of the model we obtain. We will then create test sets using one varying feature and the others remaining the same, and use the logistic regression model equation to find the probabilities of each feature of having a sleep disorder and compare them with our coefficients.

Proposed Evaluation: We can use our logistic regression model to create visualizations of our case studies using the equation, $\frac{e^{kx+b}}{1+e^{kx+b}}$, to plot the logistic curve. These curves should reflect the coefficients we obtain from fitting the model on our lifestyle data set, which will show the strength of our model in showing probabilities of having a sleep disorder.

Results:
In order to simplify and better evaluate our results, we set all the categorical variables to binary 1s and 0s. For example, for the feature ‘Gender’, we set ‘Male’ to 1 and ‘Female’ to 0. We considered ‘Overweight’ and ‘Obese’ as equivalent, as well as ‘Sleep Apnea’ and ‘Insomnia’. We created our logistic regression model on these new features, and found the attributes of it:

Feature	**Coefficient**
Age	0.152
Stress Level	0.551
Physical Activity	-0.304
Systolic	1.650
Heart Rate	0.186
Gender	-0.252
Weight	-1.180

Intercept: -0.793

From this information, we can see that the higher your systolic pressure is, the more likely you are to have a sleep disorder. We also see that if you are normal weight, it significantly decreases your likeliness of having a sleep disorder.
To evaluate these results, we carried out a case study where we created a test set that models different people with changing systolic blood pressures, keeping every other feature constant at a reasonable value. Here are the values we used in our test set:

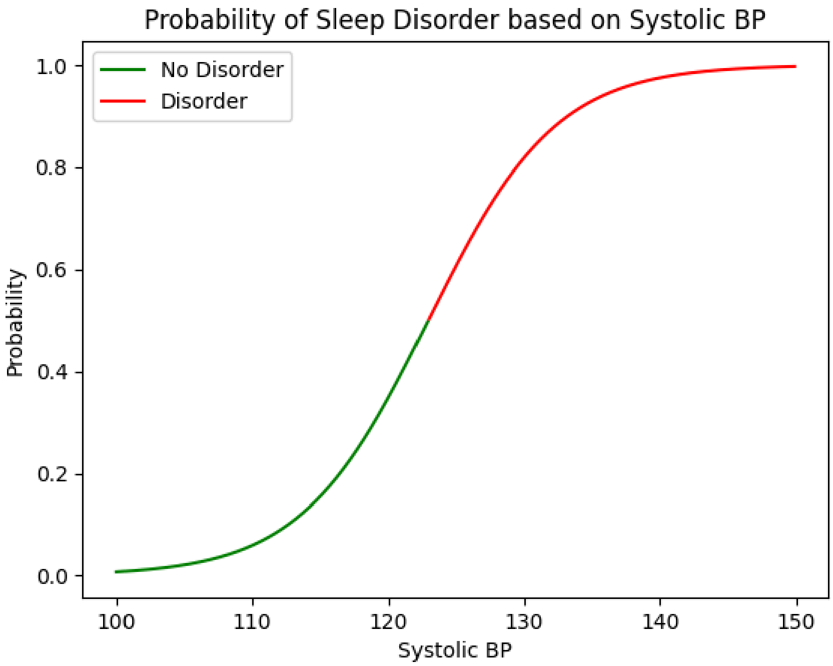
# of test cases	500
Systolic BP	**Range from 100-150**
Age	40
Stress Level	5
Physical Activity	5
Heart Rate	75

Gender	1 (Male)
Weight	0 (Normal)

Using our model, we found the probabilities of having a sleep disorder for each of these cases using the coefficients and intercept we found previously and the model equation:

$$\frac{e^{kx+b}}{1+e^{kx+b}}$$

Using this model, we created this visualization:



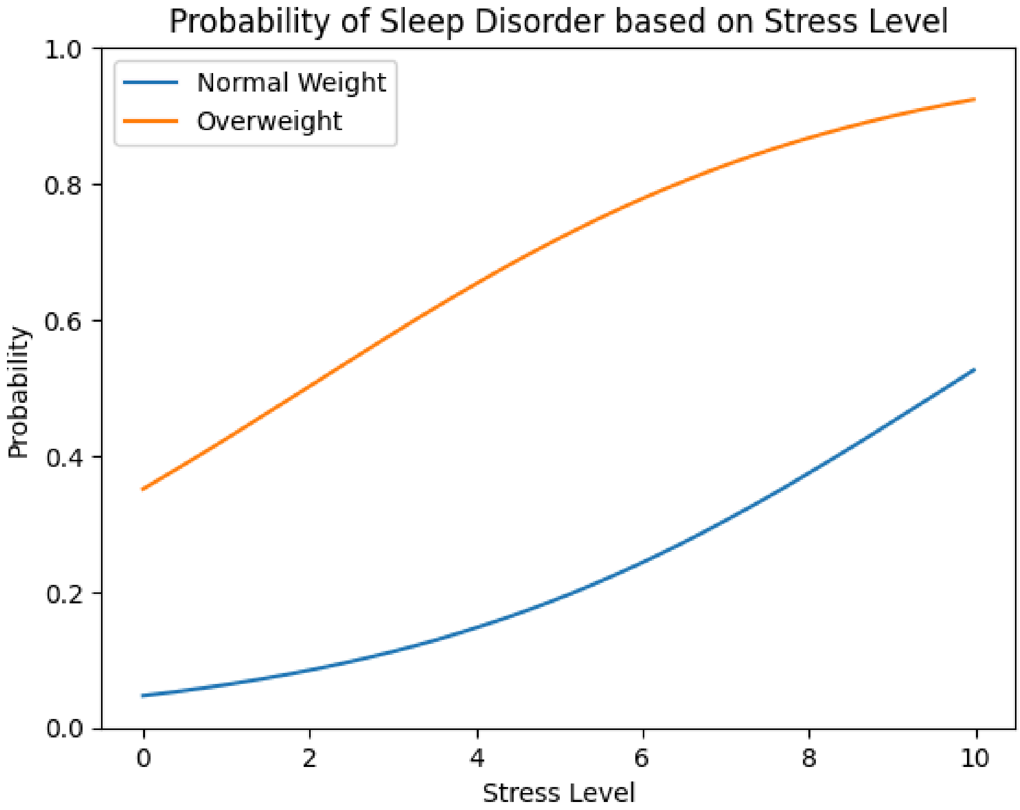
This appropriately reflects the results of our model, with systolic blood pressure having a negative coefficient of high magnitude.

We carried out a second case study, evaluating stress level and weight instead of systolic blood pressure:

# of test cases	500
Stress Level	**Range from 0-10**
Age	40
Physical Activity	5

Systolic BP	125
Heart Rate	75
Gender	0 (Female)
Weight	**1 and 0 (Normal and Overweight)**

Using the same process, we created this visualization:



This appropriately reflects our model coefficients, because a higher stress level increases your likelihood of having a sleep disorder, and being normal weight decreases your likelihood of having a sleep disorder.

Data Set 2 Question 1:

Analytical Question:

Understanding the data: can we calculate a numeric sleep score using a formula that takes into account the sleep duration, efficiency, and proportions of sleep in each stage (light, deep, REM)?

Comments:

Here are the basic statistics of the total hours of sleep required based on age group (obtained from Healthline: https://www.healthline.com/health/how-much-deep-sleep-do-you-need):

Age Group	Total Hours of Sleep Required
10 - 12	10.5
13 - 18	9
18+	7

The website also gives us this information: "All the stages of sleep are necessary, and none is better than any other. You need a balance of around 25% REM and 25% of the deepest NREM sleep to maintain your health and wellbeing."

$(sleep_{target})$ is measured from the table based on age.

$(REM_{target}) = 0.25 \times sleep_{target}$

$(deep_{target}) = 0.25 \times sleep_{target}$

$(sleep_{actual}) = duration \times efficiency$

$(REM_{actual}) = sleep_{actual} \times \frac{REM\,\%}{100}$

$(deep_{actual}) = sleep_{actual} \times \frac{deep\,\%}{100}$

Data Used:

The data we use for this question is the "Sleep Efficiency Dataset", specifically the sleep duration, sleep efficiency, the percentages for each sleep stage, and finally gender. We create multiple new variables of data using different calculations with all of these variables as well as information from Healthline.

Proposed Solution:

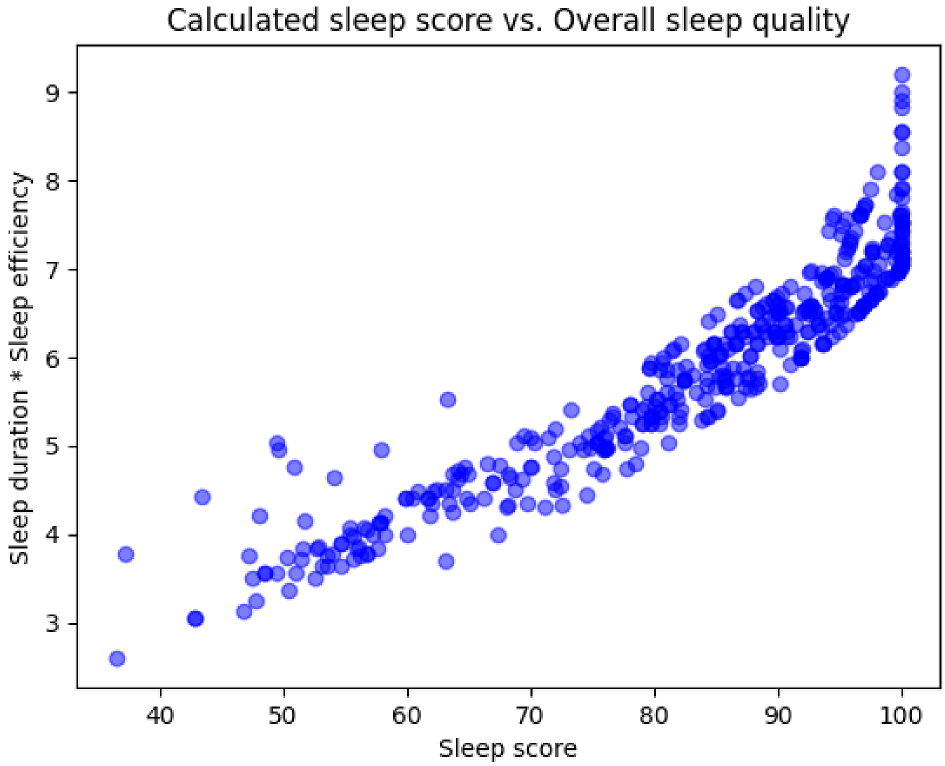
This is the formula for the sleep score that we came up with based on these formulas and the information from Healthline:

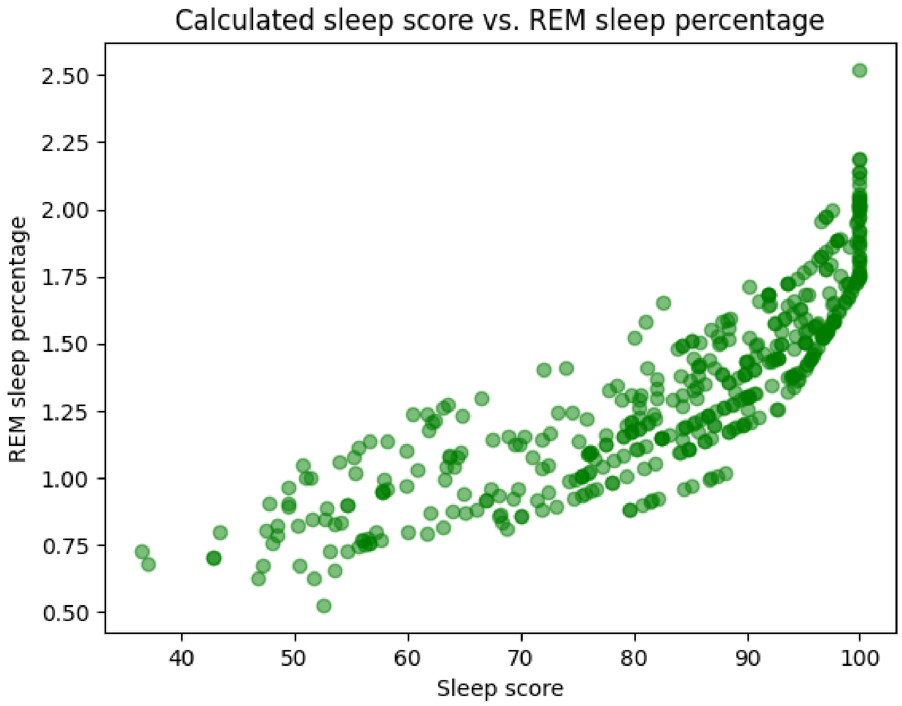
$$score = min(\frac{sleep_{actual}}{sleep_{target}}, 1) \times 50 + min(\frac{REM_{actual}}{REM_{target}}, 1) \times 25 + min(\frac{deep_{actual}}{deep_{target}}, 1) \times 25$$

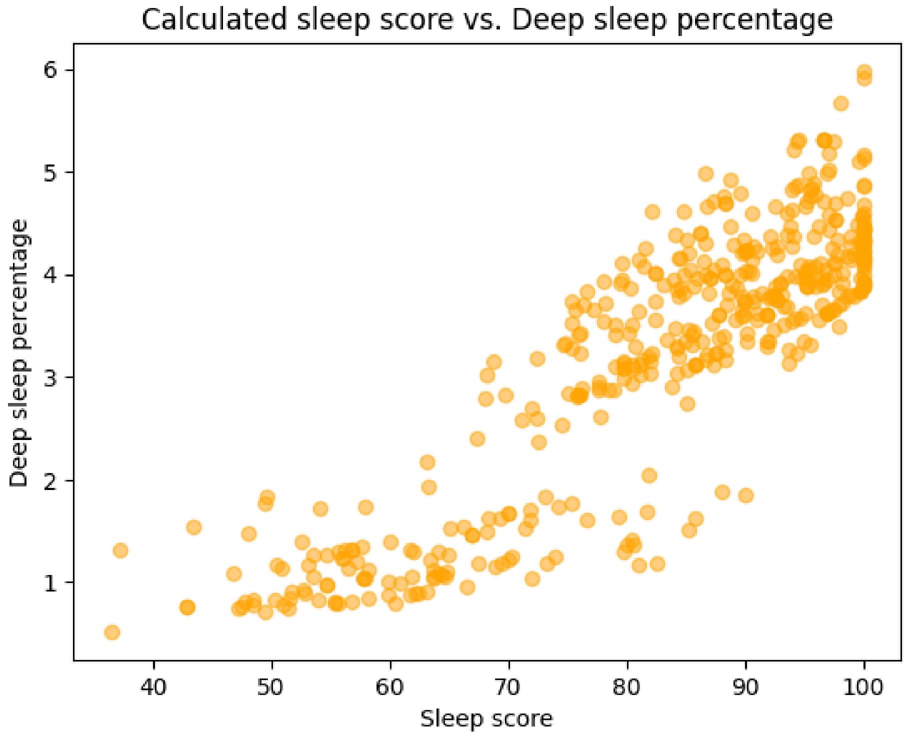
This score returns a value between 0 and 100. It gives a base of 50 points for reaching the target sleep and 25 points each for reaching both the REM target and deep target. Since there is no necessary benefit for getting more than the target values, we cap the maximum using a minimum value function between the proportion and 1.
We will implement all the formulas in code, adding columns to the dataset, and outputting a final sleep score for each entry. We will look at the correlation of this sleep score with the sleep pattern features to make sure the score is following patterns of sleep (e.g. low REM = low score).

Proposed evaluation:
The values we obtain from the formula are going to be used in future questions. We find it to be the best solution to the question of which sleep measurement should be used to carry out our predictions, as there is no concrete formula to calculate a sleep score using this specific data.

Results:
Correlations between sleep scores and Overall, REM, and Deep sleep:




Calculated sleep score vs. REM sleep percentage


Calculated sleep score vs. Deep sleep percentage

Our calculated sleep score has a good correlation with the overall sleep quality, REM sleep percentage, and the deep sleep percentage. The range of our sleep score is from the mid-30s to 100, and most of the data points have a sleep score between 50 and 100. Based on the correlation between the sleep score and all three data values, and based on the spread of the different sleep score values, this seems like a pretty good metric that takes all different stages of sleep into account.

Data Set 2 Question 2:

Analytical Question:
What are the relationships between a person's characteristics (age, gender) and sleep score (calculated in the previous question), and is there a way to predict a person's sleep score based on their age and gender?

Comments:
In the dataset used in this question, gender refers to biological sex, and the column is renamed in the Python notebook. Although predicting a sleep score by just age and biological sex without taking lifestyle habits into account probably isn't very accurate, it is still interesting to see the correlation between these attributes and how well an individual sleeps.

Data Used:
The Sleep Efficiency Dataset from Kaggle has 452 rows, with each row representing the data of an individual. It has 15 columns, six of which are used to calculate the sleep score in the previous question (sleep duration, efficiency, deep sleep percentage, light sleep percentage, REM sleep percentage, and number of awakenings). Two additional columns will be used for this question: age and gender.

Proposed Solution:
There are a few steps that we will undertake to solve this problem.

a. We will first plot all points onto a scatter plot using matplotlib.pyplot, using a different color for each gender, with the x-axis being age and the y-axis being sleep score to visualize the correlation.

b. Next, in order to try and predict a person's sleep score based on their age and gender, we will use two different regression models and compare them to each other.
    i. First model: K nearest neighbors regressor with k values ranging from 1 to 35 and other hyperparameters tuned.
    ii. Second model: Linear regression with hyperparameters tuned.

Proposed evaluation:
We will conduct a train-test split on the dataset with 20% test and 80% train. Then, we will use grid search with 10 fold cross validation to find the best model/hyperparameters for K nearest neighbors regression and linear regression. The

error of the best model from each grid search will be reported, and then the best model out of these two will be used to predict the test set. Finally, the error of the best model overall between K nearest neighbors and linear regression will be reported.

Results:
As seen in the EDA, there does not seem to be a strong correlation between age and the sleep score we calculated, which probably means that the model will not be very accurate.

RMSE values:

	Train RMSE	Validation RMSE
KNN	14.889008323155526	14.74183640335684
Linear regression	15.019596656354343	15.087918005069941
Baseline (Sample Mean)	—	15.26828515073961

Linear Regression model coefficients

Feature	Coefficient
standardscaler__Age	1.274442
onehotencoder__Sex_Female	0.538415
onehotencoder__Sex_Male	-0.538415

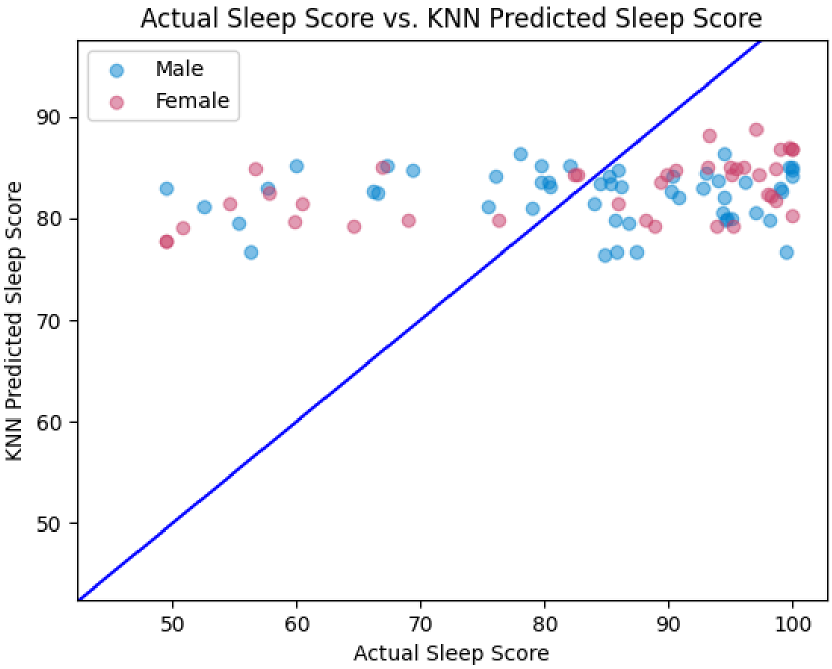
Intercept: 82.59219512572486

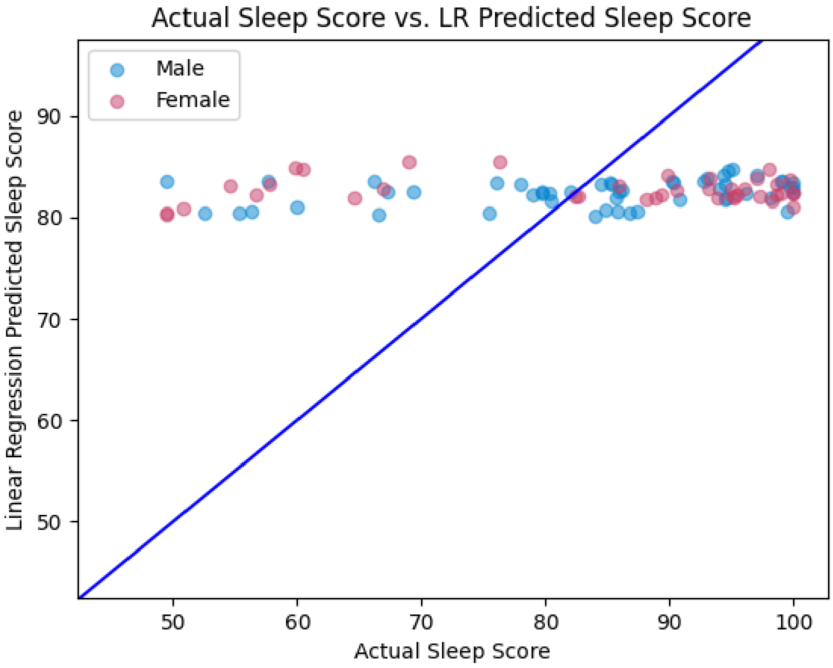
These errors are not great because the sleep score generally ranges from around 50-100. With RMSE values at around 15, it shows that the predictors' errors are spanning 30% of the actual range. The KNN model had both a lower train error and a lower validation error.

Based on the coefficients from the linear regression model, we can see that in general, a person who is older has a better sleep score, because its coefficient is positive. However, this correlation is not a very strong relationship given the variability in the data. We can also see from these coefficients that females generally have a higher sleep score than males, because the coefficient for 'Female' is positive and the coefficient for Male is negative.

We can also see here that the baseline RMSE is comparable to that of both the KNN and Linear Regression RMSE, showing that the features, age and gender, are not adding much value to the prediction, and hence, the model is not very useful in predicting sleep score.

Graphs of predictions by the models:





We can see from these graphs of the predictions that both the KNN and the Linear Regression model aren't very good at predicting the sleep score. They are both pretty much predicting a "flat" area, with a very small range of values. The KNN predictor seems to have a larger range, but it is still a very bad predictor.

It makes sense that these predictors are not very good, because as we saw in the EDA for this question, there is pretty much no correlation between age, gender, and sleep score. Looking at the graphs, and seeing that the actual range of the sleep scores in the test set is only from about 50 to 100, the validation errors don't seem as good anymore, because a RMSE of 15 would be around 30%.

Data Set 2 Question 3:

Analytical Question:

Can we make a model predicting the sleep score based on one's lifestyle (caffeine, alcohol, smoking, exercise)? Additionally, combining this analysis with the analysis from data set 1 (question 3), can we make a combined suggestion to change one's lifestyle to improve sleep habits?

Comments:

Although it is obvious that one should decrease caffeine and alcohol consumption, reduce smoking, and increase exercise to help improve their sleep score, this question is more about how much they should change those values in order to get their sleep score to increase to a "good" number. Additionally, the combined analysis with the analysis with data set 1 won't be very detailed, because there are no shared columns between the two data sets, and because the people who are included in the two datasets are also different.

Data Used:

The Sleep Efficiency Dataset from Kaggle has 452 rows, with each row representing the data of an individual. It has 15 columns, six of which are used to calculate the sleep score (sleep duration, efficiency, deep sleep percentage, light sleep percentage, REM sleep percentage, and number of awakenings). Four additional columns will be used for this question: Caffeine consumption, Alcohol consumption, Smoking status, and Exercise frequency.

Proposed Solution:

a. To predict the sleep score based on one's lifestyle, we will test two regression models and compare them to each other.
    i. First model: K nearest neighbors regressor with k values ranging from 1 to 35 and other hyperparameters tuned.
    ii. Second model: Linear regression with hyperparameters tuned.
b. To make suggestions on how to improve sleep habits, we will use these models that we have trained to find the optimal numbers for lifestyle (possibly testing different values and graphing them). Then, for a specific person, we can suggest how they can change their alcohol/caffeine consumption, exercise, and smoking habits to possibly improve their sleep habits.
c. To make a combined suggestion that will help improve sleep habits, we will look at overall suggestion trends for each lifestyle attribute (eg. decrease in caffeine

consumption, increase in daily exercise), and give an overall suggestion for what should be increased and what should be decreased.

Proposed evaluation:
We will conduct a train-test split on the data with 80% train and 20% test, and then we will use grid search with 10-fold cross validation to find the best model for K nearest neighbors regression and Linear regression. The errors of these models from the cross validation will be reported. The error of the best model out of these two will also be reported. Next, this best model will be used to give suggestions for better sleep habits. We can test the accuracy of these suggestions by finding the person in the dataset whose data is closest to the data after suggestions are applied, and comparing their sleep scores.

Results:

RMSE values:

	Train RMSE	Validation RMSE
KNN	13.189752266117278	12.979761789712958
Linear regression	13.491874254211712	13.958574299020096
Baseline (Sample Mean)	–	15.26828515073961

Linear Regression model coefficients

Feature	Coefficient
standardscaler__Caffeine consumption	0.524776
standardscaler__Alcohol consumption	-4.827315
standardscaler_Exercise frequency	2.813750
onehotencoder__Smoking status_No	3.869669
onehotencoder__Smoking status_Yes	-3.869669

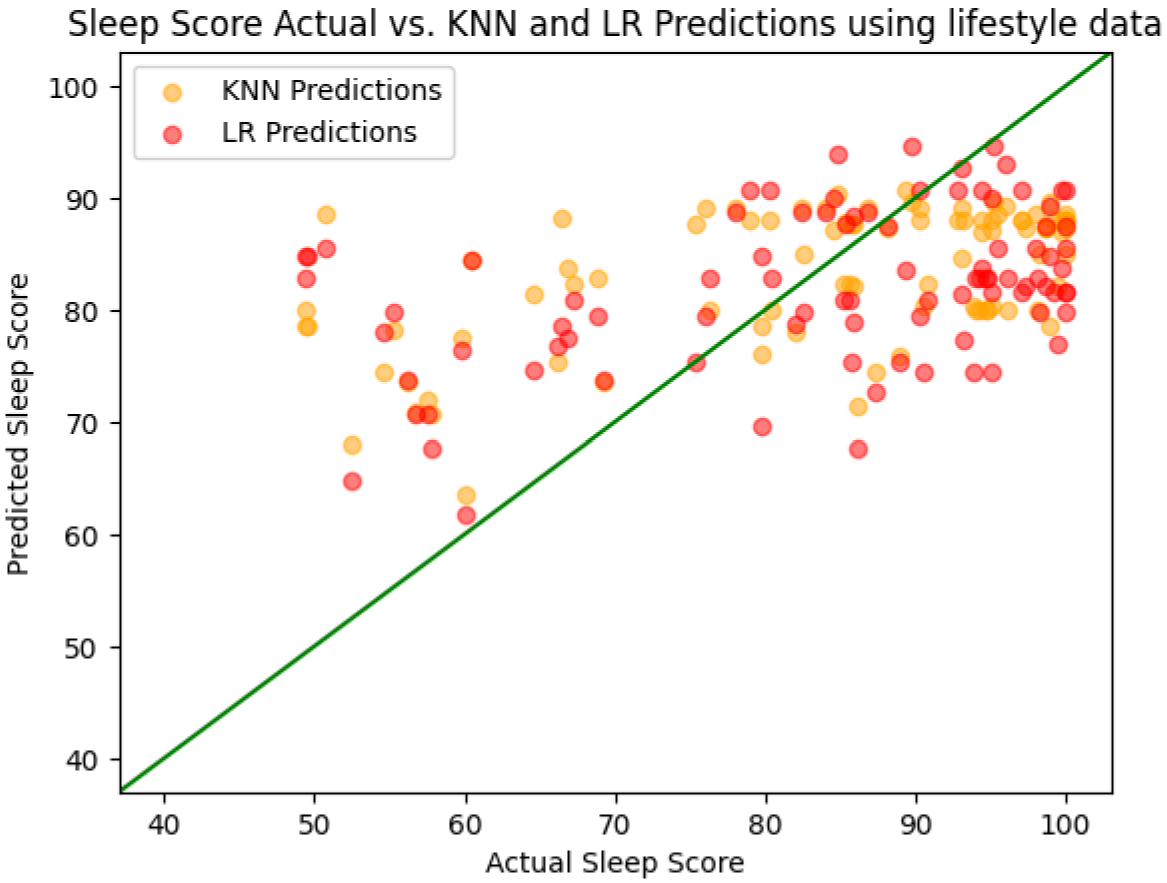
Intercept: 81.17995078356878

Using the lifestyle habits data, the KNN model also has a lower train and validation RMSE than the linear regression model. The validation RMSE is, again, lower than the train RMSE, but this is likely due to the same reason as the previous problem - the range of the actual sleep scores in the test dataset is quite small - because the train and test datasets are split the same way.

Looking at the coefficients of the linear regression model, the signs of all of the coefficients make sense, except for the Caffeine consumption coefficient. We expected this coefficient to be negative, because consuming caffeine usually results in worse sleep. It makes sense that this value is smaller than the other values, because this column has a range of 0-200 compared to 0-5 for alcohol consumption and exercise frequency, and 0-1 for smoking status. We can see from the numbers that alcohol consumption has a stronger correlation with sleep score than exercise frequency because the absolute value of its coefficient is larger.

The baseline validation RMSE here is higher than both the models' validation RMSE here again, and it is higher by more than it was for question 2, which means that the models here trained on lifestyle data are better than the models trained on characteristics.

Graph of KNN and Linear Regression predictions:

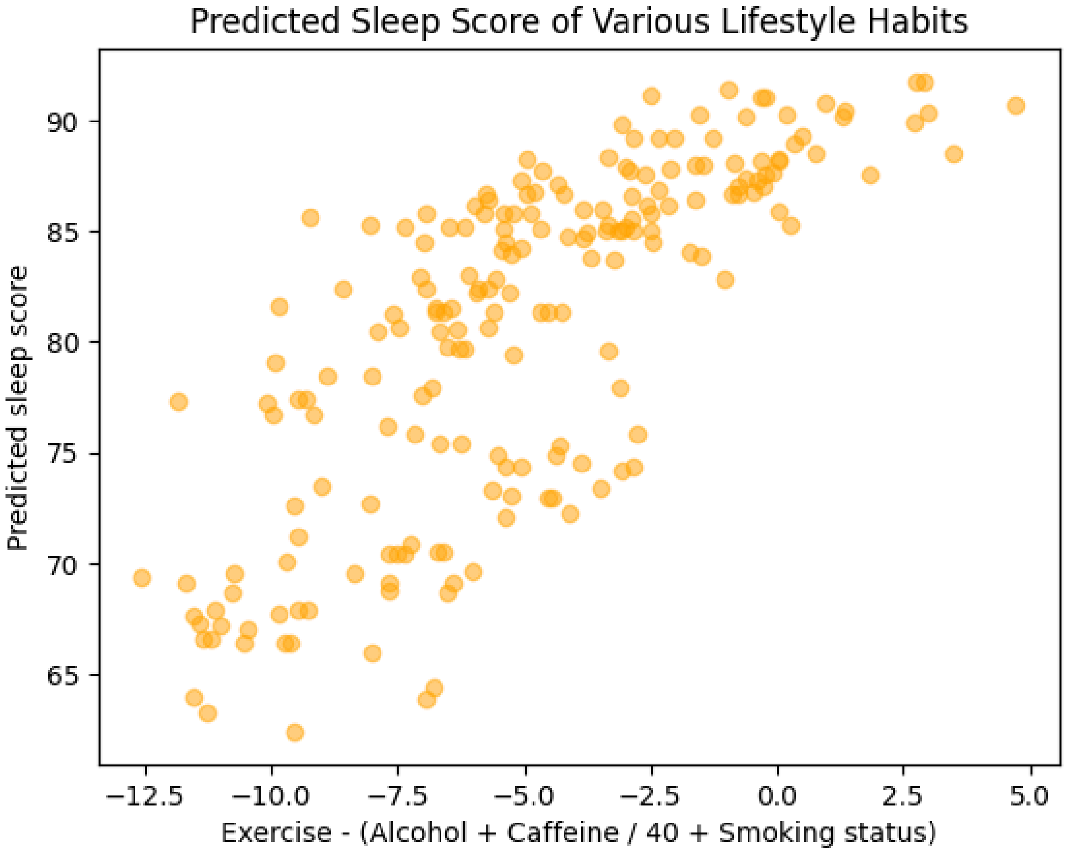


From this graph, we can see that the KNN and linear regression models which take lifestyle habits into account instead of characteristics seem to be more accurate than the models from the previous problem. They seem to be predicting a larger range of the values and their predictions are closer to the line $y = x$ overall. The range of the sleep scores calculated by both models are now from ~48-80 instead of mostly being around 70. However, the models still aren't very good, and their validation RMSEs have not gone down by a lot.

It makes sense that these models trained on lifestyle data are more accurate than the models trained on characteristics, because as we saw in the EDA, there was at least a slight correlation between the data values and the sleep score. We thought that these models would improve the RMSE by much more than they did, but this was based on the expectation that these variables had a strong correlation with the sleep score. However, again, as we saw in the EDA, the correlations were not as strong as we expected.

The graph below is from a random sample of 200 data points from a dataframe that contains all possible combinations of lifestyle values, with an x value of $Exercise - Alcohol - Caffeine/40 - 5$ (if smoking). The reason caffeine is divided by 40 is because its range is from 0-200, compared to 0-5 for most other variables.

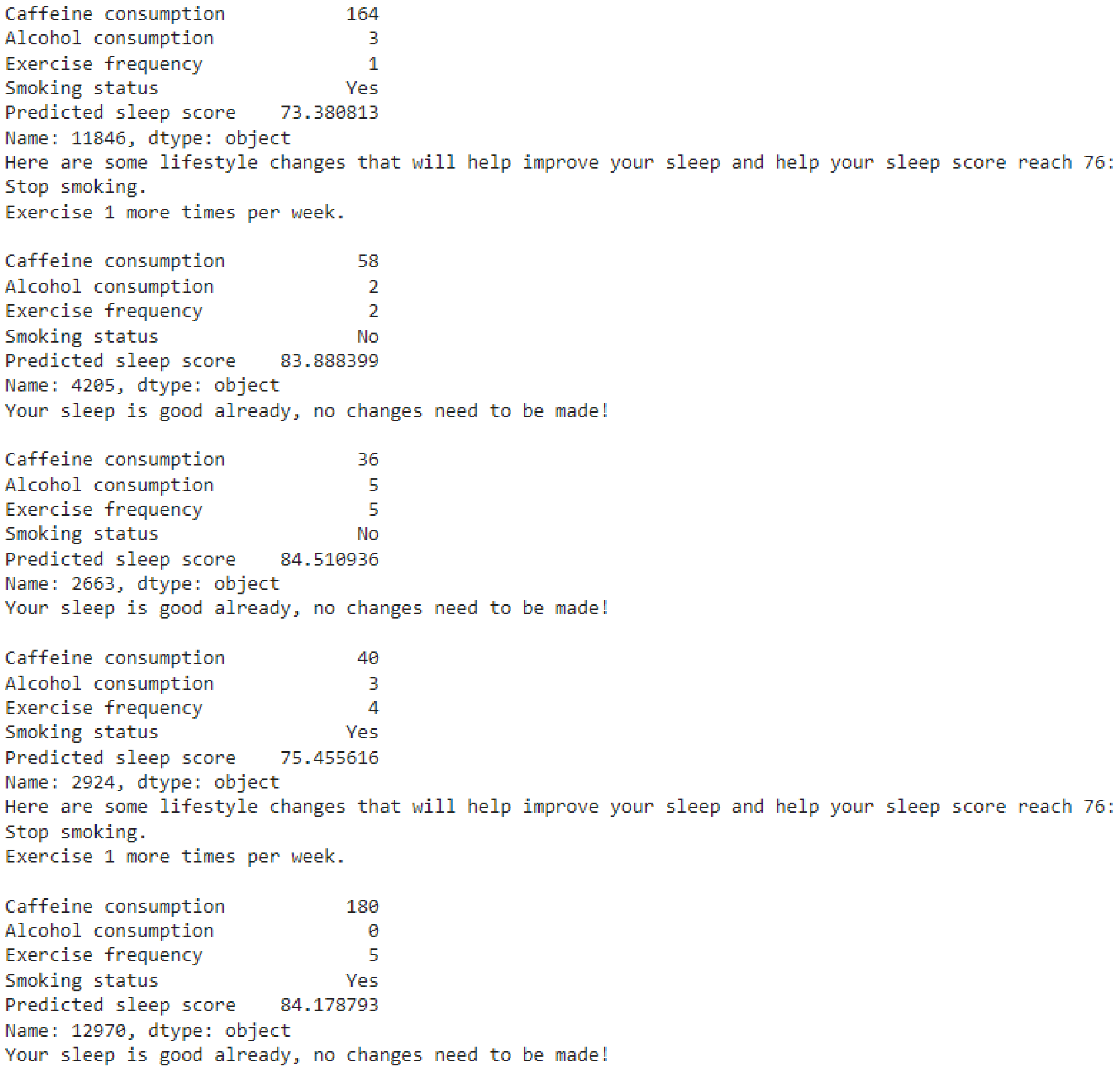
Smoking status is multiplied by 5 for the same reason.



We can see that higher alcohol consumption, caffeine consumption, and smoking results in a lower predicted sleep score and that a higher exercise frequency results in a higher sleep score.

The function we came up with to give lifestyle suggestions has two parameters. The first parameter is the data for the person, and the second parameter is the desired sleep score. The function first checks if the person smokes and suggests them to stop smoking. If their sleep score is still too low, then it decrements alcohol and caffeine consumption, and increments exercise frequency. It keeps doing this until the sleep score reaches the goal.

Here are some of the suggestions given by the function (sample of 5 randomly selected rows from the dataframe):



```
Caffeine consumption           164
Alcohol consumption              3
Exercise frequency               1
Smoking status                 Yes
Predicted sleep score    73.380813
Name: 11846, dtype: object
Here are some lifestyle changes that will help improve your sleep and help your sleep score reach 76:
Stop smoking.
Exercise 1 more times per week.

Caffeine consumption            58
Alcohol consumption              2
Exercise frequency               2
Smoking status                  No
Predicted sleep score    83.888399
Name: 4205, dtype: object
Your sleep is good already, no changes need to be made!

Caffeine consumption            36
Alcohol consumption              5
Exercise frequency               5
Smoking status                  No
Predicted sleep score    84.510936
Name: 2663, dtype: object
Your sleep is good already, no changes need to be made!

Caffeine consumption            40
Alcohol consumption              3
Exercise frequency               4
Smoking status                 Yes
Predicted sleep score    75.455616
Name: 2924, dtype: object
Here are some lifestyle changes that will help improve your sleep and help your sleep score reach 76:
Stop smoking.
Exercise 1 more times per week.

Caffeine consumption           180
Alcohol consumption              0
Exercise frequency               5
Smoking status                 Yes
Predicted sleep score    84.178793
Name: 12970, dtype: object
Your sleep is good already, no changes need to be made!
```

Data Set 3 Question 1

Analytical Question:
What similarities emerge when student's academic and lifestyle information are grouped together? What "types" of student lifestyles are observed? Are these types meaningful? What lifestyle patterns are most common among Stanford summer students?

Comments:
The purpose of this question is to analyze the specific patterns present among Stanford students based on a small sample of Stanford Summer Session students. Understanding the sleep habits of students at Stanford is useful as it allows health authorities to understand to what degree student's sleep needs to be improved along with the specific aspects of student lifestyle that should change to improve overall sleep quality. However, since the sample of students who responded to the survey mostly consisted of data science students, this will likely introduce some bias into the data; for example, if data science is generally a more difficult class, stress levels might be higher, and vice versa if data science is generally an easier class.

Data Used:
For this question, the bespoke dataset "Stanford Summer Students Sleep Survey" will be used. Specifically, the columns that will be used as inputs for the clustering algorithm are Quality of Sleep, Stress Level, Steps, and Method of Transport.

Proposed Solution:
This solution will involve constructing and training a clustering algorithm to group students based on their reported lifestyle information. The amount of clusters k will be determined through the "elbow method" by plotting the silhouette score against the number of clusters the model fits, then finding the amount of clusters where adding more clusters does not significantly increase the silhouette score of the model. In addition, the Davies Bouldin Index will be used to supplement this analysis and verify that the right amount of clusters are being formed.
After clustering, through analyzing the features of each cluster through both algebraic calculation and visualization, the patterns from each cluster can be used to draw conclusions about Stanford students during the Summer Session.

Proposed evaluation:
Since there is no ground truth in this dataset for the analysis this question is conducting, an evaluation metric that strictly uses information from a single clustering

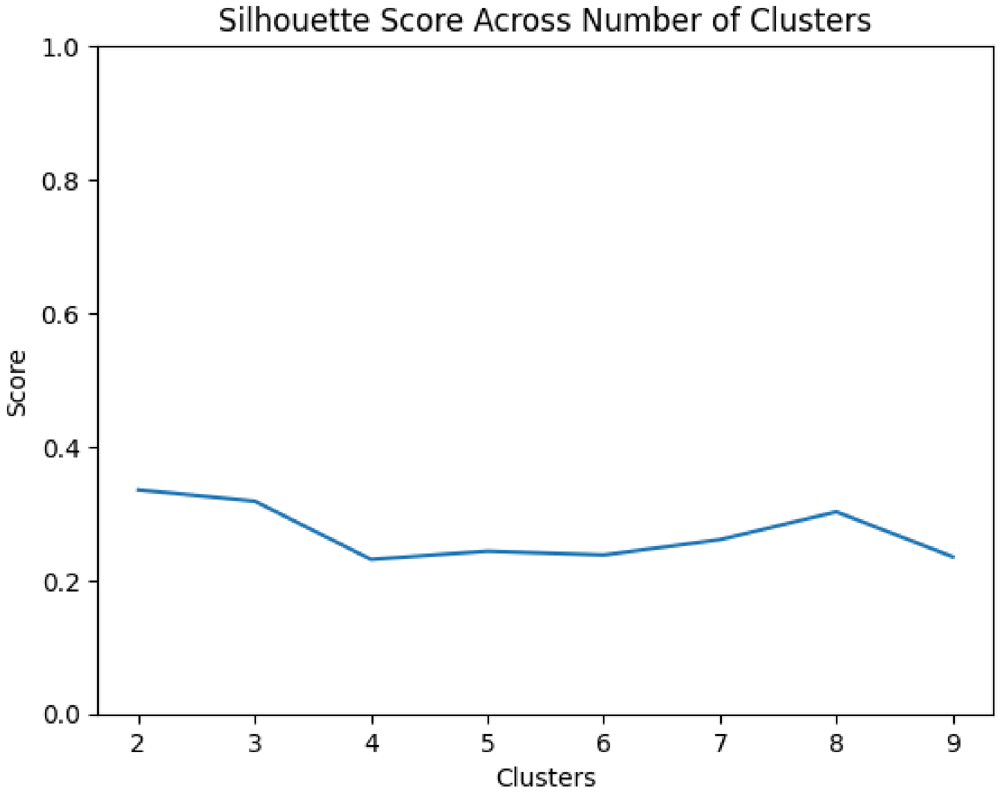
result rather than comparing two different cluster results would be the most effective. One such metric is silhouette score, which calculates the difference between the mean distance to a neighboring cluster and the intracluster mean distance. Cluster groupings for this question should attempt to maximize the silhouette score. Another such metric for verifying the silhouette score is the Davies Bouldin Index, which measures the ratio of similarity between near clusters and within the cluster itself.

Results:

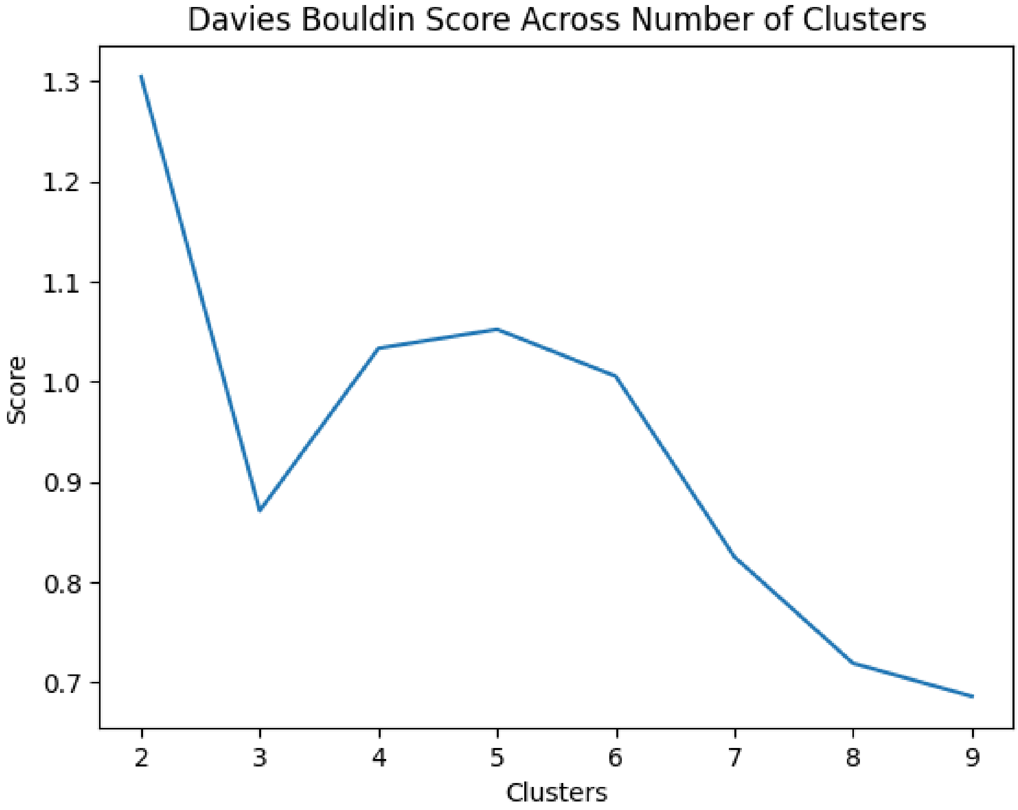
The first part of this question is transforming the independent variables being used. As all of the variables are in different scales, either subjective gradings from 1 to 10 or amount of steps in the thousands, a Standard Scaler is used to normalize the numerical features. Meanwhile, a One Hot Encoder was used to convert the categorical *Method of Transport* column into numerical features.

Next, we must evaluate the input variables, such as through comparing their correlations. When correlated columns are used to train a machine learning model, the model typically fares worse than if just one of those columns was used instead. After running a correlation check, we find that all of the correlations are within 0.4 points of 0 except for the values transformed by a One Hot Encoder. These variables are reasonable for the model.

The most important step is training the cluster model. For this analysis, we are using a KMeans algorithm. We iterate from two to nine clusters to determine the minimum reasonable amount of clusters to divide the dataset into using both the Silhouette Score Davies Bouldin Index metrics. When visualized, we get the following results:



Silhouette Score Across Number of Clusters



Davies Bouldin Score Across Number of Clusters

From these results, it is clear that the data is best grouped into three clusters. Here is the main attributes of each cluster summarized in two charts:

Means and Standard Deviations of Various Numerical Variables in 3 Cluster Model

Mean, (Std. Dev)	**Cluster 0:**	**Cluster 1:**	**Cluster 2:**
Sleep Quality	-0.13449, (0.95481)	0.06420, (1.13125)	1.14281, (N/A)
Stress Level	0.56799, (0.60057)	-1.61054, (0.57721)	-0.21160, (N/A)
Steps	-0.31281, (0.42692)	0.27844, (0.65394)	4.27279, (N/A)
Count	19	6	1

Spread of Categorical Variable *Method of Transport* in 3 Cluster Model

	Cluster 0:	**Cluster 1:**	**Cluster 2:**
Walk	6	0	0
Bike	10	4	1
Drive	3	2	0
Total	19	6	1

Overall, the clusters that form from this set of collected data are slightly weaker than a reasonable set of clusters as the *Silhouette Score* indicates an existing but slightly weak relationship while the *Davies Bouldin Index* also indicates an existing but slightly weak relationship among clusters, even in the case where the best cluster grouping with three clusters was used. As a result, care should be taken not to extrapolate information from this analysis as a certainty. One example of the weakness of this data is in the third cluster, cluster two, which only consists of a single student. This means that that specific student is an anomaly in the data, likely as their step count is incredibly high compared to the rest of the group: they have about 4.27 after standardization while the next runner-up only has 1.14. Another signifier of the weakness of the clusters is the high standard deviation among categories within each cluster. For most features, the standard deviation is higher than the mean, indicating a large possible range of values within each cluster and weakening the apparent groupings.

Despite this clustering weakness, the remaining clusters do still show a few compelling trends. Mainly, sleep quality, stress, and step count do differ significantly from group to group. The biggest factor in groupings with the least spread within each cluster was stress level. This signifies that the biggest two categories of Stanford lifestyles are based on the amount of stress a student will face. The split between stressed students and less stressed students is heavily in favor of the stressed students, who make up 19 of the 26 students included. Students belonging to the less stressed clusters tended to take more steps in a day and get slightly better sleep. Meanwhile, the more stressed cluster got slightly worse sleep and on average took less steps in a day. This indicates that sleep and stress are related; if a health official or professor was looking to encourage more students to become part of the low-stress cluster, suggesting healthy sleep habits like encouraged in the various other analyses in this report would likely assist in this goal.

Data Set 3 Question 2

Analytical Question:

Can a student’s quality of sleep be predicted using information about their lifestyle and classes? What information about a student has the greatest impact on the prediction, and how can these results be interpreted?

Comments:

The purpose of this question is to analyze the different factors related to sleep so suggestions on how to best improve it can be found. This question elaborates on the conclusions of the last question, which provides analysis on the amount of students who need to improve their sleep and which general factors are common in groups of students with low-quality sleep. This question develops that idea by focusing on the specific metrics that are related to sleep quality and being more specific about these metrics' impact, including whether some metrics are relevant at all during model selection.

Data Used:

This analysis uses the custom-constructed dataset “Stanford Summer Students Sleep Survey.” Specifically, the columns that will be searched as possible inputs for the final model are: Sex, Class Units, Method of Transport, Quality of Sleep, Stress Level, Anxiety, and Depression.

Proposed Solution:

This solution will involve training and selecting multiple machine learning models to predict the quality of sleep feature for each student in the Stanford Summer Session student dataset. These models will primarily differ in their method, i.e. K Nearest Neighbors (KNN) and Linear Regression, but also parameters and hyperparameters, such as which features are considered in the model, or the number of neighbors to consider and distance metric to use for the KNN model. These models will be selected using cross-validation and Mean Squared Error (MSE).

After the best KNN model and best Linear Regression model are constructed, they will be evaluated and compared to each other and the actual sleep scores in the dataset. For the linear regression model, the coefficients of each included variable will be considered.

Proposed evaluation:

All models produced for this question will produce quantitative predictions for the Quality of Sleep feature. As a result, the best models for each method (KNN and

Regression) will be evaluated and visualized through a train-test split with Root Mean Squared Error (RMSE), which is equivalent to the square root of the MSE. RMSE was selected as it considers the difference between a numerical prediction and the actual numerical value, but in a way that is both differentiable (unlike Mean Absolute Error, MAE) and understood in linear units (unlike MSE). Generally, models that minimize RMSE are considered to be more valuable.

Results:

First, the columns that will be searched are selected. Then, model selection takes place. For the linear regression model, we conducted a search using Pandas and a for loop to iterate through each possible subset of features using itertools. Then, for the KNN model, a combination of both Pandas and Sci-kit Learn is used; the features are selected through the same Pandas methodology as the linear regression model, but within the each feature subset, the best hyperparameters such as k_neighbors are selected with a GridSearchCV fit-transform method.

Attributes of Best Models for Each Regression Method

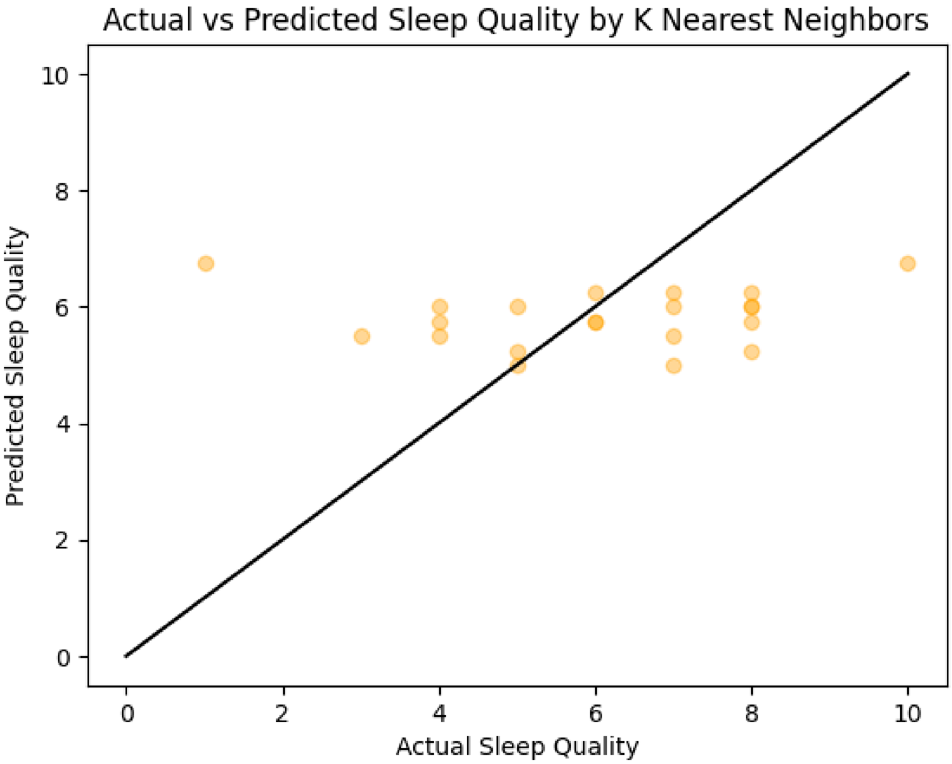
Method	**Linear Regression**	**K Nearest Neighbors**
Best Independent Variable Features	Sex, Stress Level	Sex, Method of Transport, Stress Level, Anxiety
RMSE (in CV searches)	1.794	1.814
Hyperparameters	N/A	Metric = Euclidean, n_neighbors = 4

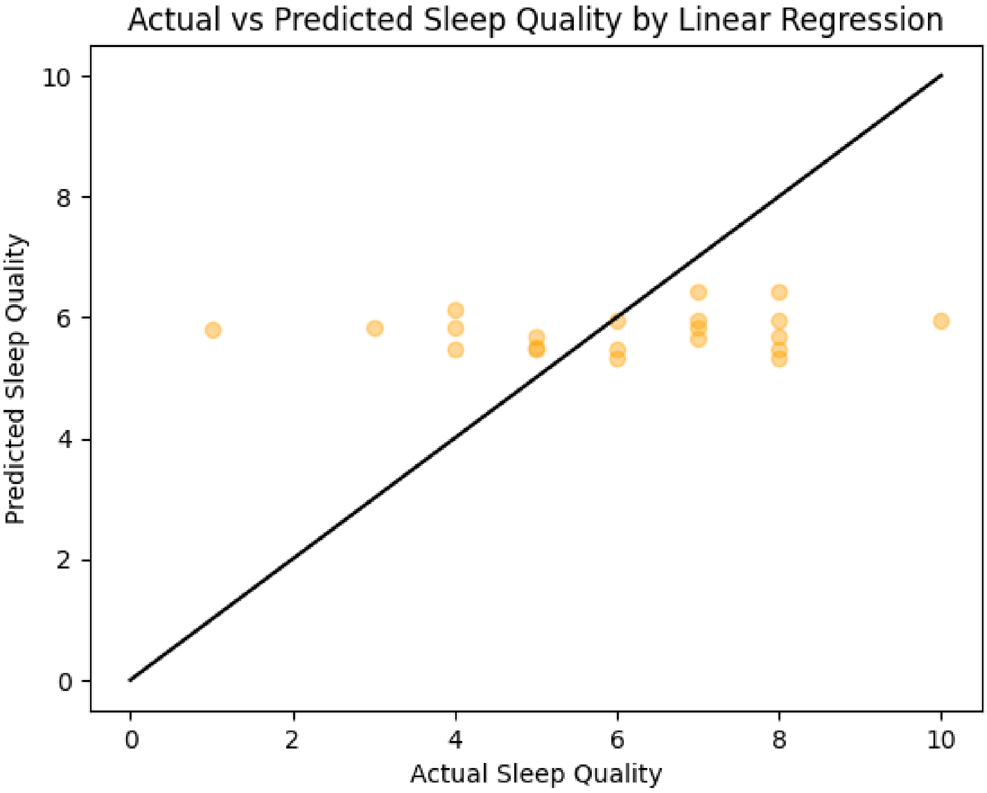
However, when these models are compared to a simple model that only returns the mean of its training with a train-test-split protocol, the results showed that the models were unreliable and worse than the simple mean method.

RMSEs of Various Models on Train-Test-Split Data

Linear Regression	**K Nearest Neighbors**	**Mean**
4.2481	4.3780	4.2358

A train-test split ratio of 0.5 testing was used as there are very few rows of this dataset. This may play some part in the abysmal test results. Mainly, each model did not have much data to train on. In addition, similar to the results from Data Set 2 Question 2, it appears that there is no correlation between any of the features of the dataset and the quality of sleep each student reported. To gain more insight, the predictions were graphed over the actual values:





From these visualizations, it is clear that the predictors are not reliant on their independent variables to make predictions. In the K Nearest Neighbors model, it is clear that the neighbors of each test case student are typically representative of the mean of the entire dataset, meaning that the data is generally centered around the mean. Interestingly, this pattern does not apply at the extremes of sleep quality, which appear to have more in common with each other than more moderate ratings.

In the linear model, all of the coefficients are close to zero. This indicates that there is not much variation in the model across variable values, which signifies that the searched features are generally poor predictors of sleep quality.

Ultimately, the Stanford dataset suffers from being short and lacking data overall, along with a poor selection of features. Future datasets should be constructed to be more comprehensive and to range from a wide array of sources.

Discussions and Conclusion

In total, we looked at three datasets. Dataset 1, “Sleep, Health, and Lifestyle” looked at sleep quality for professionals with different occupations. Dataset 3, “Stanford Summer Session Sleep Survey” looked at similar data for high school and university students. Dataset 2, “Sleep Efficiency” looked at the quality of sleep based on more definitive factors such as the phases of the sleep cycle and lifestyle choices. We analyzed different features such as characteristics, lifestyles, and vitals in order to predict quality of sleep as well as sleep disorders.

The primary purpose of this study was to identify the significant factors impacting general sleep health and habits in order to make recommendations for certain lifestyle changes and health choices.

In the first dataset, we saw that stress levels had a significant negative impact on quality of sleep for both males and females. This trend was also observed in data set 3 amongst Stanford students. This is a significant finding from our study because it shows how minimizing stress would result in improved sleep quality.

Another important factor identified from data set 1 was body weight. Overweight and obese patients had poorer sleep quality than that of normal weight patients, and were at a higher risk of suffering from sleep disorders such as Sleep Apnea and Insomnia.

A third critical factor that was identified was blood pressure. Having a normal blood pressure, both systolic and diastolic, improved sleep quality and reduced chances of suffering from a sleep disorder. Specific case studies demonstrated significant reduction in the probability of having sleep disorders. In one case study, we found that changes in your blood pressure could result in a decrease from a 100% probability to a 0% probability of having a sleep disorder.

The 3rd case study looked at lifestyle habits, and though it is obvious you will get benefits, the models predicted how much your lifestyle should change to get desired sleep improvements. It can also be used as a motivational tool to help patients quit smoking, reduce drinking, and exercise more frequently. Compared to these factors, caffeine consumption had a lesser effect on sleep quality.

The models developed in this study could be individualized for patients to see possibilities to improve their sleep quality by making certain changes, and can be used as guidance for physicians to aid and motivate their patients to make these changes in their lifestyle and health. It can also be used as an indication to prescribe medicine to control something like a patient’s blood pressure. Further, patients predicted to have high probabilities of sleep disorders can take preventative or coping measures such as using a CPAP machine to prevent Sleep Apnea.

As seen in Dataset 2, age and gender did not have much of an impact on sleep habits. This just shows that there is no “right time” to start improving your health and getting better sleep. Everyone has an equal opportunity to improve their sleep.

The overall limitations of this study is that the three different data sets were used to make this analysis; a singular, larger data set would have been more definitive. The first two datasets had numbers of entries in the mid 100s, with the “Sleep, Health, and Lifestyle” data set having around 370, and the “Sleep Efficiency” dataset having around 450, while the smaller Stanford student data set we created had around 50 entries. A much larger dataset with all the features covered throughout the three would definitely improve the quality of predictions and results. Nonetheless, the analysis here has been significant in identifying ways for improving lifestyle to positively impact quality of sleep, and provides a useful resource for service care providers and individuals who want to look out for their health.