Sleep Health in College Students: A Multivariable Predictive Modeling Analysis

Jonah Watson

Department of Computer Science

Adrian College

Supervisor Dr. Yasser Alginahi

A capstone project proposal submitted to the Department of Computer Science in partial fulfillment of the requirements for the Degree of Arts in Data Analytics at Adrian College.

Fall 2025

Abstract

The purpose of this study was to investigate the factors that contribute to sleep duration among American college students. This study utilized data from the American College Health Association's National College Health Assessment III (ACHA-NCHA III) dataset. Data cleaning and analysis were performed using Python and the Matplotlib Pyplot, NumPy, Pandas, Seaborn, Scikit-Learn, and Statsmodels libraries. A logistic regression model was developed to predict optimal weeknight sleep (7-9 hours) versus suboptimal weeknight sleep (less than 6 hours). Psychological, diagnostic, behavioral, and trauma-related predictors were incorporated into the model. The model achieved an accuracy score of 0.615 and a ROC-AUC score of 0.637, indicating moderate predictive power. Stress was identified as the strongest negative predictor, with a 26% decrease in the odds of optimal sleep for each increase in stress levels. Overall health rating was the strongest positive predictor, with a 22% increase in the odds of optimal sleep for each increase in overall health rating. Diagnosed insomnia also showed a strong negative association. Unexpectedly, the presence of anxiety increased the odds of optimal sleep. The analysis also revealed significant comorbidity between insomnia and other conditions such as ADHD, anxiety, depression, sleep apnea, and substance use disorder. These findings reinforce the established connection between mental health and sleep quality documented in prior literature and highlight the multi-dimensional nature of the factors contributing to sleep quality in college students.

Table of Contents

1 Introduction

Sleep is an essential aspect of life that affects all areas of health, and yet is often overlooked by the average person. Across college campuses worldwide, students often face many late nights. Some of these late nights are spent studying or completing homework, some are spent sharing quality time with friends, and others are spent partying hard. As a result, sleep quality frequently suffers. Mornings often begin with energy drinks, lectures are met with tired eyes, and students' bodies and minds are unable to perform at optimal capacity. In several studies, nearly 60% of college students are classified as having poor sleep [1]. Among these students, approximately 15% experience academic difficulties, along with a multitude of other negative health and cognitive effects [2]. Contributors to poor sleep are varied and can include lifestyle and behavioral factors such as poor diet and sedentary lifestyle [3], loneliness and psychological distress [4], high levels of stress [5], alcohol consumption and substance abuse [6], and caffeine intake [7]. Evening screen use can also interfere with sleep [8], as can traumatic experiences [9]. On the other hand, the consequences of poor sleep are poor academic performance [3], excessive daytime sleepiness [10], impaired learning and memory capabilities, increased mood disorders such as depression and anxiety, and risky behaviors such as substance misuse and drowsy driving [7].

The goals of this research are two-fold. First, it aims to raise awareness about sleep and expand the readers' understanding of its effects by investigating what factors contribute to sleep quality at the university level. Sleep is critical for individuals of all ages, and at the college level, this is especially true, as many students balance early classes, late nights, athletic commitments, and a substantial homework load. Although most people recognize that sleep is important, fewer are aware of the more nuanced negative consequences of poor sleep quality and what contributes to it. Another objective is to contribute new information to the topic. Conducting this research and presenting the findings will help increase awareness, and any new discovery or additional support for current theories can advance the field of research. College is a large undertaking for many people financially, mentally and emotionally. If college students understand the importance of prioritizing quality sleep, their lives will benefit in many ways, and they will be able to get as much out of college as they can.

This study will focus on behaviors and consequences that are associated with both adequate and poor sleep quality, using very recent data taken from a large sample of around 100,000 US college students from hundreds of institutions. There have been many previous studies that have investigated the relation between sleep quality and various contributing factors and consequences, so this study will serve as a deeper dive into many of these elements and will also attempt to shed light on less discussed factors, such as certain mental health conditions and experiences, with a large sample and a credible source to support the findings. The goal is not to remove the foundation set in place by previous studies, but to add support to the ideas that already exist and are well researched, as well as to draw attention to less discussed factors so that future research can continue to account for many elements relevant to sleep. In doing this, both academic institutions and individuals can receive well-researched recommendations through this study, to ideally reduce sleep-related problems in the future. In the next section, relevant studies on this topic will be discussed in more detail.

2 Literature Review

The importance of sleep is widely known and agreed upon, but some nuance is often lost. People understand that they need to get quality sleep for their health, but what are the true consequences of poor sleep quality at the college level? Most people would agree that the primary goal of going to college is to develop the soft and hard skills relevant to your field of interest and then undertake the process of finding a job which exercises these skills in some way. Learning in itself is one of the most fundamental parts of college: learning and growing skills that are relevant to your field of study, learning how to apply for jobs and brand yourself in general, and learning other important life lessons. Poor sleep quality greatly inhibits these abilities, and unfortunately, many college students do not get the most out of their sleep. This review of the literature will cover the factors that contribute to poor sleep quality in college students, especially behaviorally, as well as the negative consequences that arise as a result of poor sleep. Clinically defined sleep disorders, such as insomnia, will be discussed in this review of the literature along with other disorders and mental health hurdles. Poor academic performance will also be investigated as an outcome of poor sleep quality. Physical activity, sedentary behaviors, caffeine use, alcohol use, and screen time are contributing factors that will be investigated here as well, most of which have negative effects on sleep quality.

A comprehensive study by Zisapel, [8], laid a good foundation for the rest of this review of the literature. The study covers the role of melatonin in human sleep and the regulation of circadian rhythms. Humans have a central circadian clock, found in the suprachiasmatic nucleus (SCN), a small region in the hypothalamus in the brain. This 'clock' synchronizes physiological rhythms with the 24-hour day and night cycle, and interacts with melatonin in different ways accordingly. Light suppresses melatonin production, whereas in the absence of light, melatonin release conveys a darkness signal to the body and SCN. This melatonin production is what drives the desire to sleep, and this urge usually begins to occur around 2 hours after melatonin production begins. Melatonin shifts the physiological state of human bodies towards sleep in a natural and non-sedative way, unlike the sedative nature of drugs. Melatonin activation also interacts with the Default Mode Network (DMN), signaling it to reduce mind-wandering and contribute to the desire for rest. Certain diseases and disorders suffer from reduced or delayed melatonin production, which correlates with non-restorative sleep, disrupted circadian rhythms, and worsening of health. Having a consistent wake-up routine with limited use of light exposure in the evening and healthy daily behaviors contributes to restful sleep and robust circadian rhythms. In cases where melatonin production is misaligned, such as travel and circadian rhythm sleep disorders, clinical trials show that supplementation with melatonin can help the body return to a more natural circadian state. This study was conducted primarily on animal models and small human trials, so the findings cannot be directly correlated with college students. However, basic principles have been well established: maintaining a consistent sleep schedule and participating in healthy daily behaviors and activities, such as physical activity and limiting light in the evening, are strongly related to healthy sleep quality, with the option of supplementing melatonin and possibly also seeking the advice of various care specialists if problems arise or disorders are present [8].

Liu et al., [11], conducted a study investigating the relations between poor sleep quality and its related risk factors among university students. This study included 1317 university students in China, with 64.6% being female. This study used a cross-sectional survey that utilized the Pittsburgh Sleep Quality Index (PSQI) to assess sleep quality. Statistical tests such as logistic regression and the chi square test were performed on the data. From this sample of students, 30.1% students had poor sleep quality according to the PSQI (a score greater than 5). Physical activity, both low and high intensity, was identified as a strong association with better sleep quality, while demographics such as sex and income were not significantly associated in most cases. In this study, taking naps more than 3 times a week was associated with restful sleep, and the authors say that they believe that more research should be done on how physical activity and napping behaviors influence sleep quality. There is also evidence that longer sleep duration, better sleep quality, greater sleep consistency, and their correlations with better academic performance are especially apparent days before final exams or other assessments. However, this study is limited in scope, being limited to a single university [11].

The study conducted by Phan et al., [12], utilized a small sample size of 39 first-year college students, but made compelling findings in the realm of sleep quality and memory capacity in college students. This study was observational, with students reporting daily in a journal about their duration and quality of sleep, and participating in working memory tasks. The researchers compared their daily sleep patterns with their performance on memory tasks. Students who slept fewer hours per night showed significantly worse performance in memory tasks and also reduced working memory capacity [12], which is the limited amount of information that a person can store and manipulate in their mind in short intervals around 15 seconds, critical for learning and academic success [13]. There was also a stronger correlation between consistent daily sleep patterns and memory performance than there was for total daily sleep duration and memory performance. This study, however, is limited by its small sample size of 138 students, and is also slightly weakened by not accounting for caffeine use and screen time, which, in fairness, is something lacking from much research on this topic [12].

Since caffeine use and screen time are often not the main factors in this area of study, it is important to find a specialized study or studies on these topics. The study by Hershner and Chervin investigates the causes and consequences of sleepiness in college students and covers the topics of caffeine use and screen time, among others. This study was a narrative review on US college students that brought together evidence on sleep patterns, prevalence of sleepiness, and contributing factors to sleep issues such as caffeine use, alcohol use, and screen time. Some key findings from this comprehensive literature review were that a large proportion of college students do not meet recommended sleep durations, and many report frequent daytime sleepiness. Caffeine and energy drinks are widely considered a coping mechanism for sleepiness and academic demands, but also contribute to the sleep problems themselves, such as delayed sleep onset, fragmented sleep, and shortened total sleep time. The use of screens and artificial light has been well documented to suppress melatonin and therefore inhibit sleep quality and circadian rhythm [8], and this study also found exactly that. Screen time in the evening was also shown to be pervasive and cause issues such as delaying circadian rhythm timing (the body's internal clock), reducing melatonin secretion, and reducing sleep duration in general. The consequences of poor sleep identified in this article are lower GPA, impaired learning and memory capabilities, increased mood disorders such as depression and anxiety, and risky behaviors such as substance use and drowsy driving. Although this study does not inherently provide new data, it is a comprehensive review drawing from nearly 100 sources and discussing a few more niche factors that contribute to poor sleep [7].

In a more specific study, Patrick et al. investigated how caffeinated energy drink use and binge drinking predicted the sleep quantity, quality, and tiredness of college students. This study utilized a sample of 667 US college students who utilized web-based daily reports for four semesters, up to 56 consecutive days per semester, resulting in around 25,000 days of data. The students reported whether they consumed energy drinks or binge-drank on a given day and also reported their sleep quality, quantity, and tiredness that night. On days when students reported drinking an energy drink, they also reported sleeping significantly fewer hours, reporting a worse subjective sleep quality, and feeling more tired the next day. In general, students who tended to drink more energy drinks also tended to have poor sleep outcomes, and the outcomes were even more negative for those students who tended to binge drink. Ultimately, this article gives credit to the idea that caffeine, and more specifically energy drinks, as well as binge drinking, have significant negative effects on sleep health [14]

A study conducted by Peltzer and Pengpid, [6], covered nocturnal sleep problems in detail, covering 20,822 college students from 26 countries. The students represented countries in Asia, Africa, and the Americas and were evaluated using self-reported questionnaires, where the data were interpreted with logistic regression to identify predictors of sleep problems and also reported by country. Students were asked questions about sleep, such as difficulty initiating sleep, maintaining sleep, unplanned early morning awakening, and non-restorative sleep, as well as questions about health behaviors, such as smoking, alcohol consumption, physical activity, and depressive symptoms. Of the 20,822 students, around 10.4% had frequent difficulty initiating sleep, 8.3% had difficulty maintaining sleep, and 7.1% had difficulty waking up in the morning. Higher rates of sleep problems were associated with mental health problems, substance use, and poor academic performance. There was considerable variation in the total percent of students with sleep problems reported in each country, with women consistently reporting more sleep problems than men. Some possible limitations of this study, however, are that sleep data was self-reported and not measured by any machine, and also that, due to the cross-sectional nature of the survey, causal conclusions are not quite fair to draw in this case [6].

Sleep problems have also been reported in much higher ratios among smaller sample sizes and more specific locations. This fact, combined with the varying data by country in the Peltzer study, raises questions about how location can play a role in sleep quality. One of these two studies was conducted by Mahfouz et al. at Jazan University in Saudi Arabia. This study investigated the levels of sleep quality and physical activity of 440 students. The students were sampled with cluster random sampling, and the study used a cross-sectional observational design. Sleep quality was measured using the Pittsburgh Sleep Quality Index (PSQI), physical activity was measured using the International Physical Activity Questionnaire (IPAQ), and mental health indicators were measured using DASS-21. Of the 440 students, 63.9% reported poor sleep quality, and 62.7% of the students were physically inactive. Physical activity was significantly associated with better sleep quality, and students who were more physically active were also more likely to have good sleep quality, with an adjusted-odds ratio of 1.72, a moderate association. A primary factor limiting this study, however, was that all samples were taken from one university in Saudi Arabia and performed cross-sectionally [15].

Another aspect of sleep and its relation to the other aspects of the lives of college students is its interplay as both a contributor and a result of academic stresses, irregular schedules, and sleep patterns, as discussed in the study by Deliens et al. This study was conducted with a small sample size of 46 students in Belgium but sheds more light on sleep's relations with other aspects of students' lives. In this study, students were chosen from a variety of disciplines and placed into seven focus groups led by a semi-structured question guide. The primary topics included sleep routines and daily routines and how they interact with physical activity or lack thereof. The timing and quality of sleep were shown to affect students' energy levels and their motivation for physical activity. This sedentary behavior was also shown to be more prominent during times of higher academic stress, such as exam periods and times where schedule irregularities existed. This study demonstrates that a certain level of self-discipline is required to balance quality sleep and productive physical activity with the stresses that come with being a college student. However, because the data are self-reported and reported within a focus group setting, some bias may be present, which may have limited this study, in addition to the small sample size. Both the participants and the authors recommended the idea of promoting sports and other on-campus activities to greater effect and lowering the barrier for participation in some cases to increase the reaches of these programs and develop the physical health of the students while lowering the amount of sedentary time they spend. [5].

Russell et al., [16], conducted a study investigating the correlations between mental health and sleeping behaviors of college students, especially in terms of self-harm and suicide. This study was a systematic review that used more than 90 different studies from university/college students around the world. From this multitude of studies, Russell et al. found that sleep problems (defined here as difficulty falling asleep, staying asleep, and non-restorative sleep) are reliably associated with an increase in suicidal and self-harming behaviors. In addition, these studies established the understanding that sleep disturbance is an independent risk factor and is an important contributor to negative health outcomes. Nightmares and insomnia symptoms were also found to have a connection to the risk of suicide in students. However, in using so many studies, the constituents of a "sleep problem," for the sake of this study, are likely to be a bit opaque as different studies use different measures to define poor sleep.. This study greatly reinforces the connection between sleep quality and self-harming behaviors and demonstrates the need for psychological intervention and the promotion of maintaining a healthy headspace [16].

In terms of recent research on the topic, the study by Chung et al. is among the newest. This study utilized a sample of 798 American college students aged 18-24 and covered the topics of substance use, psychological distress, and loneliness, and their role related to academic impairment from sleep difficulties. The data came from the spring 2021 American College Health Association National College Health Assessment (ACHA-NCHA III) dataset, very similar to the ACHA-NCHA IIIb dataset that is used in the firsthand analysis in this paper. The ACHA administers the NCHA each spring and fall to American college students as part of an ongoing national study. Students from hundreds of institutions respond to the survey, and it is the largest resource on student health, with more than 2.5 million students in the database historically. The survey contains upwards of 800 variables covering all aspects of college student health, and the ACHA makes public executive summaries and downloadable reference-group reports available. Data were analyzed using logistic regression in an attempt to measure whether psychological distress, loneliness, and risky substance use were associated with academic impairment due to sleep difficulties, where demographics were control variables. In this model, demographic variables such as sex and year in school showed a weak correlation, whereas many of the other variables showed a strong correlation. There was a significant association with variables such as loneliness, psychological distress, high-risk scores for alcohol and marijuana, and their respective correlations with academic impairment due to sleep difficulties. Despite alcohol and marijuana use showing such a strong correlation, tobacco showed a noticeably weaker correlation. With psychological distress and social isolation being such widespread issues and sleep and academic performance being important for college students, the authors suggest initiatives to eliminate isolation and reduce the potential for self-harming behavior [4].

Continuing with recent research, and research that uses ACHA-NCHA survey data, is a study by Lederer et al. This study had a large sample size of 39,146 American undergraduates from 75 institutes, with the data being collected before COVID-19 in this case. Lederer et al. performed a secondary analysis of the ACHA-NCHA III data to analyze health-related behaviors and academic achievement among college students. Although sleep was not the main focus of this study, there is still relevant information here, and the large sample size is a notable factor in the findings. Self-reported GPA and 33 health behaviors were investigated with methods such as weighted cross-tabulations and multinomial logistic regressions. The authors found that many health-related behaviors were significantly associated with academic performance. Among these is sleep behavior, which turned out to be a strong predictor in several comparisons. Students with overall better health behaviors (including better sleep) tended to have higher GPAs. The differences between A GPA students and D GPA students demonstrated in this study were significantly different across many categories, most notably sleep quality, with 61.4% of A GPA students getting more than 7 hours of sleep per night, compared to just 40.5% of students with a D GPA. The disparity is also observed in diet, physical activity, sedentary behavior, sexual risk behavior, and mental health [3].

Another study with a robust sample size is the study conducted by Upright et al. This is a generalized study on health issues and their effects on the academic performance of college students, with around 20,000 students sampled from a college in the southern United States. These students completed the ACHA-NCHA II assessment, and the results are descriptive, as the students selected which health issues negatively affected their academic performance, and then the frequencies of each option were reported. The second most reported issue was sleep difficulty, with 15.5% reporting that it affected their academic performance, only behind stress, with 20.8% of students reporting it. Rounding out the top ten were work at 12.7%, anxiety at 12.0%, cold/flu/sore throat at 11.9%, relationship difficulties at 10.4%, internet use/computer games at 10.1%, depression at 8.8%, alcohol use at 8.4%, and death of, or concern for, friend/family member at around 7%. It is important to note that many of these factors that contribute to poor academic performance also contribute to poor sleep quality, as proven by many of the other articles in this review of the literature, most notably anxiety, depression, and alcohol use. This study gives further confirmation to the idea that poor mental health, poor academic performance, and poor sleep quality are interwoven. However, a minor issue with this study in relation to sleep is that all sleep difficulty reports are self-reported and do not necessarily all have a clinical diagnosis as a backbone [2].

Dinis and Bragança, [13], conducted a study focused specifically on depression and sleep in young adults and college students. This research effort includes 32 studies that utilize a college student/young adult population and directly focus on the connection between sleep and depression. The most common evaluation tools among the studies were the PSQI to measure sleep quality and the Center for Epidemiologic Studies Depression Scale (CES-D), to measure depression symptoms, used in 22 and 13 studies, respectively. Of the 32 included studies, 13 used a North American population, 10 used an Asian population, and the rest were a relatively even spread between Oceania, Europe, the Middle East, Africa, and multi-regional. The sample sizes included range from less than 50 all the way into the upper hundreds, although not every sample size is mentioned. In this review of the literature, there is a consistent association between poor sleep quality and depressive symptoms in these populations, both as a contributor to poor sleep quality and also as a result of poor sleep quality. One study investigated that poor sleep can make it more difficult to disconnect attention from negative stimuli, increasing the prevalence of depression symptoms. This study revealed that depression and sleep have a bidirectional relationship, as they both influence each other's behavior. Sleep quality seems to be more consistently a predictor of depressive factors than the opposite, however. This study proves yet again that improving sleep quality can help both prevent and relieve depressive symptoms in college students [17].

ADHD is another mental illness that correlates with poor sleep quality. A recent study by Demirkan et al., [18], focused specifically on this relationship, investigating a sample of 503 Turkish college students to attempt to connect the dots between elevated ADHD symptoms and a range of sleep disturbances. The metrics utilized were the Dream Anxiety Scale, PSQI, and Epworth Sleepiness Scale (ESS) for sleep, and the Beck Depression Inventory (BDI) for evaluating depression symptoms. The study found that students with elevated ADHD symptoms had significantly worse sleep, indicated by higher PSQI scores, greater daytime sleepiness, indicated by higher ESS scores, more depressive symptoms, and higher dream/anxiety disturbance. This same group also had more reports of nightmares, more frequent sleep disruptions, and were more likely to sleep less than recommended. The study included regression models where the sleep and mood variables together explained around 68% of the severity of the ADHD symptoms, indicating that there is a strong overlap between ADHD, mood, and sleep disturbances. Students with ADHD traits would benefit from interventions targeted specifically for them, as ADHD symptoms are connected to a myriad of sleep disturbances and mood difficulties, and interventions geared more towards anxiety or depression may not appropriately address the same issues experienced by those with elevated ADHD [18].

Continuing with investigating the links between mental health conditions and sleep quality in college student populations is the study by Kim et al., [19]. This study explores how sleep patterns and sleep quality are related to mental health issues, particularly ADHD symptoms and depression. This cross-sectional study was conducted with a group of 290 South Korean university students, with a mean age of 22 years. Sleep quality was assessed with the PSQI, and sleep patterns were observed through bedtime, wake time, and sleep duration, and the variability of each aspect. In terms of sleep quality, more extreme and undesirable ADHD and depression symptoms were significantly correlated with poorer sleep quality. For sleep patterns, greater variability in sleep was associated with poorer mental health measures. Kim et al. also concluded that sleep quality remained a significant predictor of depressive symptoms with covariates being accounted for. This study sheds light on the symptoms of ADHD as another mental health issue correlated with poor sleep, along with depression. Regulating sleep patterns to maintain sleep consistency and using targeted interventions can help college students avoid psychological distress and improve mental quality-of-life. However, a limitation to be noted with this study is the relatively small sample size and regional-only data. [19].

Another study investigating the link between sleep quality and depression was conducted by Wang et al., [20]. This study also focuses on anxiety related to sleep and offers some interesting insights. This study used a sample of 4,325 college students from two colleges in China's Tibet region. Data was collected using an online questionnaire, and common metrics such as the PSQI were utilized to measure sleep quality, the Generalized Anxiety Disorder Scale (GAD-7) for anxiety symptoms, and the Patient Health Questionnaire (PHQ-9) for depressive symptoms. Another resource, titled the Cognitive Emotion Regulation Questionnaire (CERQ) was used to investigate adaptive and maladaptive strategies related to these issues. The study showed that out of the 4,325 students, around 46% had poor sleep quality, 37% exhibited anxiety symptoms, and 52% exhibited depressive symptoms. Similar to several of the other studies included in this review of the relevant literature, poorer sleep quality was significantly associated with a higher severity of anxiety and depression, especially with sleep being a predictor of these issues, rather than being a result of them. In terms of Cognitive Emotion Regulation Strategies (CERS), adaptive CERS such as acceptance and positive refocusing showed a small effect, and this was limited to the path of depression influenced by sleep. In contrast, when students negatively regulated their emotions or regulated their emotions maladaptively, as labeled by the CERS, poorer sleep quality was more strongly correlated with psychological distress. This indicates that students who ruminate on negativity, catastrophize, and blame themselves for their negative emotions exhibited a considerable association with poor sleep quality as well as the exacerbation of anxiety and depression symptoms. This helps shed light on how poor sleep quality contributes to anxiety and depression. Overall, this study contributed to the evidence found in many other studies that poor sleep quality is strongly associated with elevated anxiety and depressive symptoms among college students. Maladaptive CERS have a stronger mediation than adaptive CERS, but this mediation is undesirable, as it exacerbates the symptoms of depression and anxiety. This suggests that an effective strategy to improve quality of life in this area is like a chain reaction: improve sleep quality, experience fewer symptoms of anxiety and depression, and reinforce this by replacing the stronger maladaptive CERS with adaptive CERS and, in general, making an effort to keep a positive head space. College students would benefit not only from interventions that focus on sleep quality, but also from emotional regulation support, so that these negative CERS can be more easily avoided. Similar to a handful of other selected studies, a primary weakness lies in the data gathering being performed in one specific geographical and cultural area. The sample size in this study, however, is much more substantial in comparison to other studies that also have a geographically-specific scope [20].

The sleep difficulties of college students can also be informed by trauma and other adverse childhood experiences (ACEs) such as maltreatment and other household dysfunction. The study conducted by Albers et al., [9], investigated the connection between ACEs and sleep difficulties among young adult college students. This study was conducted with a sample size of 4,013 students from Texas and California and had a predominantly female sample size of 70.8%, with logistic regression being the primary statistical analysis method. The key findings of this study include that around 40% of students report having at least one ACE, and for that group, the chances of having sleep difficulties were much higher than those without ACEs. The chances were as high as 2.4 times more likely for someone who endured maltreatment during their childhood to have poor sleep quality than for someone who had not. For those who experienced more than one ACE, there was an even greater chance of having insomnia or other sleep disorders. ACEs are also very strongly related to contributing to mental health issues such as depression and anxiety, and can lead to other issues such as substance abuse, further reinforcing the correlations between poor sleep quality and its contributors. There is likely some recall bias present with this study, as people were asked to revisit their memories of childhood trauma and self-report it, but the study shows strong evidence that ACEs are connected to poor sleep quality and poor academic performance [9].

The study by Becker et al., [1], more intentionally views the connections between males and their sleep as well as females and their sleep, and also gives more support to the idea that mental health and sleep are strongly correlated. This survey involved 2,890 college students from seven different U.S. universities, and used self-report surveys to obtain data on mental health and sleep problems, where correlation analyses were used to measure the validity of variable relationships. A key finding from this article, among others, is that roughly 27% of students met the criteria of having at least one sleep problem, and among these problems were widespread insomnia symptoms, poor sleep quality, and irregular sleep schedules. Inside of the study, there is also nuance in comparisons between male and female sleep quality. Women in this study were more likely to report insomnia and poor sleep quality, and reported taking longer to fall asleep than men, using sleep medications more often than men as well. Men reported slightly more sleep apnea symptoms, but overall reported having better sleep, with 64% of females meeting the established PSQI value to classify them as poor sleepers compared to 57% of men meeting this threshold. Males also reported going to bed later and waking up later than females on average. Sleep problems in this study were also strongly associated with higher levels of depression, anxiety, and ADHD symptoms. Academics were not involved in this study, but from the other studies present in this review of the literature, there has been a strong intertwining of mental health, academic performance, and sleep quality, showing that each of these three categories often informs the other. However, in this study, similar to many other studies, sleep quality information is self-reported and not clinically diagnosed in all cases, and while findings on the surface level would likely remain the same, obtaining more information on specific sleep disorders related to this topic would help interventions narrow their focus when applicable [1].

The next two articles are more in depth on insomnia and sleep apnea and how they affect college students, according to previous studies, respectively. The first of these is a study conducted by Taylor et al., [10], that investigates insomnia in college students and how it affects mental health, quality of life, and substance use difficulties. This study used a sample of 1,039 students from the University of North Texas. Notable population statistics include 72% participants who report as female and a mean age of 20.39 years. The students involved in this cross-sectional study received a survey and a one-week sleep diary. In this study, insomnia was defined by DSM-5 criteria, where there is a complaint of insomnia greater than (or equal to) 3 nights per week for more than (or equal to) 3 months, with daytime dysfunction. 9.5% of the sample met the criteria for chronic insomnia, and interestingly, 6.5% had complaints of insomnia but did not meet the criteria, whereas, in contrast, 26.9% met the criteria but did not report any insomnia complaints. Students with chronic insomnia had poorer health in many aspects: worse sleep diary metrics (lower total sleep time, lower sleep efficiency, and more unintentional awakenings), increased mental and physical fatigue, elevated anxiety, elevated depression, an increase in perceived stress, and a lower quality of life. Surprisingly, however, the differences between the chronic insomnia group and the normal sleeper group were minimal in terms of excessive daytime sleepiness, cumulative GPA, and substance use, apart from a greater use of hypnotic and stimulant medications for the insomnia group. The insomnia group also reported more medical problems at 44.8% versus 24.8% for the normal sleeper group. Despite sharing several minor flaws with many other studies done on the general topic of sleep and college student health (causal conclusion issues due to cross-sectional survey design, self-reported data, data from one campus), this study covers many correlations between students with and without chronic insomnia and health measures [10].

The second of these specialized sleep disorder studies is the research performed by Jain and Verma, [21]. This research uses a sample of 1,524 college students, 43% male and 57% female, and was conducted at the Integral Institute of Medical Sciences and Research in Lucknow, India. The students were issued the SLEEP-50 questionnaire, which covers a number of sleep-related disorders, where the questions were based on a 4 point Likert scale. In this group, 25% of all students tested positive for at least one sleep disorder. Narcolepsy was reported most frequently with 18%, followed by obstructive sleep apnea (OSA) with 11%, circadian rhythm disorders (CRDs) with 6%, insomnia with 4%, nightmares with 3%, and sleepwalking with 1%. Some of the prevalent risk behaviors were a 17% report rate of alcohol use at night and an 11% report rate of having a noisy sleep environment. Sadness and a lack of pleasure at bedtime were also reported, but in smaller percentages. This study also found that a portion of these students with sleeping disabilities were also at academic risk, with a total of 35% for CRDs, 31% for narcolepsy, 25% for OSA, 22% for insomnia, 17% for sleepwalking, and 12% for nightmares. Although this study did not have a large sample size or information from multiple nations or institutions, the correlations in the findings are still strong enough to raise awareness of the idea that sleep disorders of all kinds negatively affect academic performance [21].

Noisy sleep environments are not a factor often addressed in studies of this type, but understanding that, logically, noisy sleep environments are typically worse for sleep quality raises questions about sleep experiences for students living on-campus in dormitories as well as fraternity and sorority housing. There have been a small number of studies done on these topics with regard to sleep, but a study done by Dockery, A.H., [22], is a recent example that investigates this. This study has a very small sample size of 74 students from Middle Tennessee State University, but the study uses the sleep hygiene index (SHI), which is defined by the American Academy of Sleep Medicine (AASM). The sample size is divided between students living on campus and off campus. However, no elements of the sample are explicitly mentioned as living in a sorority or fraternity. The students in the sample completed questionnaires that assess their SHI, which measures habits and practices considered to promote good sleep, sleep efficiency, and sleep quality through a sleep diary and questions, and anxiety as a covariate. The study showed that there were no significant differences in sleep hygiene scores between residents on campus and off campus, even after controlling for anxiety (not allowing anxiety to skew the results, as it is widely considered to be a prominent factor contributing to poor sleep). Off-campus residents, however, did have a higher sleep efficiency than on-campus residents, even when accounting for anxiety. The fact that on-campus residents have lower sleep efficiency points to the idea that they likely had more disruptions inside or spent more time in bed awake than off campus residents [22].

With all of these sources saying what contributes to and results from lack of sleep or poor sleep quality, how do we know what an optimal amount of sleep per night is on average? Perhaps the best supported research currently available is the study conducted jointly between the American Academy of Sleep Medicine and Sleep Research Society, [23]. This study utilized 5,314 scientific articles and used a 12-month process to scrutinize and perfect the findings. The current evidence shows that getting 7 or more hours of sleep per night on a regular basis promotes optimal health in adults aged 18 to 60. There is some minuteness in individual variability based on factors such as genetics, behavior, environment, and medical factors, but the authors state that generally aiming for 7 or more hours is the best decision. Sleeping less than 7 hours on a regular basis was associated with a number of adverse health outcomes, such as weight gain, depression, and hypertension, among others. Poor immune function, increased pain, impaired performance, and a greater chance of erroneous actions and accidents were also associated with consistently getting less than 7 hours of sleep. The authors also stated that sleeping more than 9 hours a night is acceptable for young adults or individuals recovering from lack of sleep or illness, but for others, it is uncertain whether sleeping 9 or more hours poses health risks, suggesting that aiming for 7-9 hours per night is a safe range [23].

This project will address several questions regarding college students and the quality of their sleep. This study will attempt to investigate similar topics prevalent in previous literature, such as correlations between sleep and mental health, sleep and academic performance, sleep and physical health, and other relevant lifestyle patterns among college students. More importantly, however, this study will also attempt to shed more light on some relevant topics that often exist in the shadow of more broad topics. These include interactions between sleep and a number of psychological, behavioral, emotional, diagnostic, and trauma-related factors. The following methodology will cover the data science pipeline that will be used as a guide for the process of developing this research report.

3 Methodology

This research study primarily serves as a secondary analysis of survey data. A multi-stage data science approach was used to examine sleep and its relationship to other behaviors and traits in students. Data used in this project belongs to the American College Health Association (ACHA) and their National College Health Assessment (NCHA) of spring 2024. The methodology consisted of many stages, beginning with data collection and profiling to assess the scope and structure of the data. Data cleaning and preprocessing were applied next to ensure that variables were prepared and fit for analysis, including the removal of duplicate entries, handling of missing values, and conversion of categorical responses to numerical codes. Next, exploratory data analysis and visualization were conducted to identify meaningful patterns, distributions, and correlations within the data. From these insights, statistical models were constructed and tested to evaluate the relationships between sleep and health outcomes. Finally, the data was polished and refined to be presentable in a clear manner, so that significant findings and implications on college health research can be presented with clarity. Figure 1 shows the methodology used in this study.

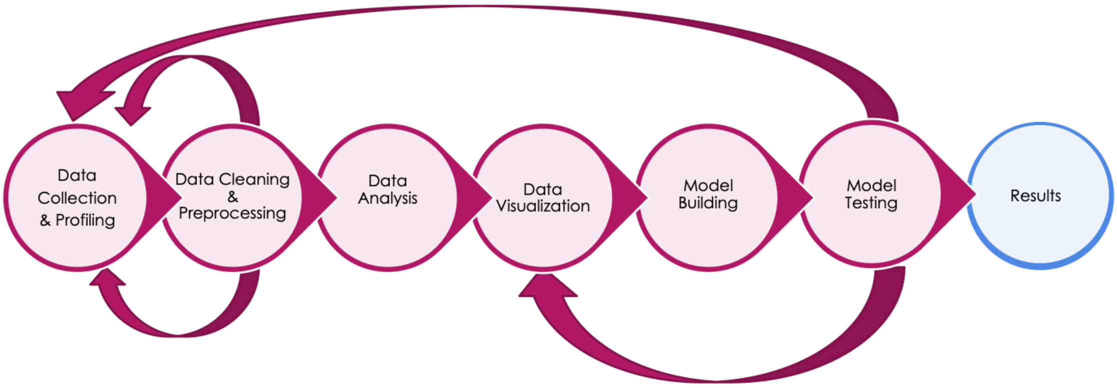


Figure 1: The process used in tandem with the data for this study.

3.1 Data Collection and Profiling

The dataset for the first-hand analysis in this study was sourced from the ACHA-NCHA survey, with this edition administered in spring 2024. This dataset includes responses from 103,639 students representing 154 American higher education institutions, making it one of the largest NCHA samples to date. The survey was administered between January and June. The dataset was received as a .SAV file type, and converted to a .CSV file for use in Microsoft Excel and VSCODE with Python. The majority of the survey questions contained preset categorical answers with correlated numeric codes. The survey included various question formats, such as variations of Likert scales, yes/no questions, and items that used numeric bins. Using the Pandas module in Python, general information about the dataset was obtained, including the number of rows and columns (.shape), metadata (.info()), summary statistics (.describe()), counts of missing values per column (.isnull().sum()), and counts of duplicate rows (.duplicated().sum()).

The results of these methods indicated that each column contained an average of around 2,000 rows with missing information. One question on the severity of chronic conditions on academic performance contained 28,067 missing rows—nearly four times as many as any other variable. This is likely to be attributed to students without chronic conditions having no reason to respond. For the purpose of this study, data from 20 survey questions and four additional metrics were analyzed, resulting in a total of 42 columns due to multipart questions. Refer to Appendix B: Profiling for the Python code used for this section of the pipeline. For clear reporting purposes, profiling results were presented before cleaning results, though in practice, cleaning was performed first so that demographic statistics and other tables and visualizations were presented untarnished by duplicates and other issues. This is why Appendix B is referenced before Appendix A here.

The variables of the ACHA-NCHA spring 2024 survey that will be used in this study are provided in Table 1. For additional details, please visit the ACHA's website at this link: `https://www.acha.org/wp-content/uploads/NCHA-IIIb_SPRING_2024_REFERENCE_GROUP_DATA_REPORT.pdf`.

Table 1: Description of Variables

Variable	**Description**
N3Q1:	Self-reported overall health rating
N3Q3E:	Hours spent doing physical activity each week
N3Q3I:	Hours spent partying each week
N3Q13:	Time it usually takes to fall asleep
N3Q14:	Average amount of sleep on weeknights in the last two weeks (excluding naps)
N3Q15:	Average amount of sleep on weekend nights in the last two weeks (excluding naps)
N3Q16A:	Days in the last 7 days when you woke up too early and could not fall back asleep
N3Q16B:	Days in the last 7 days when you felt tired or sleepy during the day
N3Q16C:	Days in the last 7 days when you had an extremely hard time falling asleep
N3Q16D:	Days in the last 7 days when you got enough sleep to feel rested
N3Q16E:	Days in the last 7 days when you took a nap
N3Q20D:	Has the student been a victim of sexual assault in the last 12 months?
N3Q20F:	Has the student been a victim of rape in the last 12 months?
N3Q20G:	Has the student been a victim of stalking during the last 12 months?
N3Q41C:	Is the student engaged and interested in their daily activities?
N3Q42B:	Does the student tend to bounce back after illness, injury, or hardship? (resilience)
N3Q48:	Overall stress level in the last 30 days
N3Q65A2:	Has the student been diagnosed with ADD/ADHD?
N3Q65A3:	Has the student been diagnosed with alcohol or drug abuse/addiction?
N3Q65A7:	Has the student been diagnosed with anxiety?
N3Q65A15:	Has the student been diagnosed with depression?
N3Q65A28:	Has the student been diagnosed with insomnia?
N3Q65A35:	Has the student been diagnosed with sleep apnea?
N3Q65Y:	How much have diagnosed conditions affected academics in the last 30 days?
N3Q66P:	In the last 12 months, how have sleep difficulties affected academic performance?
N3Q67A:	Sex
N3Q69:	Age
N3Q72:	Current year in school
N3Q75A1:	Race/ethnicity — American Indian or Native Alaskan
N3Q75A2:	Race/ethnicity — Asian or Asian American
N3Q75A3:	Race/ethnicity — Black or African American
N3Q75A4:	Race/ethnicity — Hispanic, Latina, or Latino
N3Q75A5:	Race/ethnicity — Middle Eastern, North African, or Arab
N3Q75A6:	Race/ethnicity — Native Hawaiian or Pacific Islander
N3Q75A7:	Race/ethnicity — White
N3Q75A8:	Race/ethnicity — Biracial or Multiracial
N3Q77B:	Do you live in fraternity or sorority housing?
N3Q80:	Approximate grade point average
RBMI:	Body Mass Index categories (WHO definition)
RKESSLER6:	Kessler-6 score for mental distress (0-12 negative for mental distress, 13-24 positive)
RULS3:	UCLA Loneliness Scale (3-5 negative for loneliness, 6-9 positive)
DIENER:	Diener Flourishing Scale (8–56; higher = greater well-being)

3.2 Data Cleaning and Preprocessing

Several rounds of data cleaning were performed as the focus of the research narrowed, which involved removing variables deemed less relevant or unnecessary for the analysis. Most of the survey questions included preset answers for respondents to select, each answer associated with a numerical value. For the purpose of analysis, all responses were converted from their original categorical form (e.g., very likely) to their corresponding numerical values (e.g., 5), and the columns were renamed to track what each number represented for each question. One singular duplicate row was identified and removed with the method .drop_duplicates(). Due to all of these variables used in the study being bound to a certain range, no winsorization or removal of outliers was performed. Some variables present are formed from groups of other, more specific variables, with only the summarized versions being used for the sake of this study. These include RBMI, which stands for body mass index which is derived from a person's weight and height, RKESSLER6, which includes six questions on mental health struggles and associated distress, RULS3, which includes three questions on loneliness, and DIENER, which is a scale that represents a group of questions on overall psychological well-being. Sleep apnea and insomnia are two sleep disorders that are typically considered well-known but will be defined here for the sake of clarity. Sleep apnea is a sleep disorder that causes people to stop breathing during their sleep. The brain tries to protect the body by waking the person up enough to breathe, but this prevents restful and consistently healthy sleep [24]. Insomnia is a sleep disorder that causes someone to have trouble falling asleep, staying asleep, or waking up too early. Each of these issues can affect sleep quality in its own way [25].

For the sake of making visualizations easier to follow and a bit more concise, some variable response bins were reshaped. Attention was given to avoid distorting the findings or presenting directly manipulated data. For example, despite having more concise categories than the original format, the RBMI variable was still divided using the specific thresholds between underweight, healthy weight, overweight, and obesity; however, the levels of obesity were collapsed under a single variable labeling obesity as a whole. Refer to Table 2 for the participant demographics of this study, and Appendix A: Cleaning for the Python code used for this section of the pipeline.

Table 2: Participant Demographics

Characteristic	**N (%)**	**Mean**
Study Sample Size	103,639	–
Sex		
Female	72,586 (70.9%)	–
Male	29,775 (29.1%)	–
Age	–	23.4
Year in School		
Freshman	20,832 (20.7%)	–
Sophomore	17,239 (17.1%)	–
Junior	20,600 (20.5%)	–
Senior	15,796 (15.7%)	–
5th+ Year Undergraduate	4,779 (4.7%)	–
Master's Student	13,646 (13.6%)	–
Doctorate Student	8,198 (8.1%)	–
Race/Ethnicity		
American Indian/Native Alaskan	2,354 (2.3%)	–
Asian/Asian American	17,425 (16.8%)	–
Black/African American	6,920 (6.7%)	–
Hispanic/Latina/Latino	19,120 (18.4%)	–
Middle Eastern/North African/Arab	1,985 (1.9%)	–
Native Hawaiian/Pacific Islander	615 (0.6%)	–
White	62,857 (60.6%)	–
Biracial/multiracial	5,183 (5%)	–
Approximate GPA		
A	63,718 (63.7%)	–
B	29,899 (29.9%)	–
C	5,841 (5.8%)	–
D/F	551 (0.6%)	–
Weight (RBMI)		
Underweight	4,927 (4.9%)	–
Healthy Weight	51,516 (51.8%)	–
Overweight	24,386 (24.5%)	–
Obese	18,704 (18.8%)	–
Serious Psychological Distress (KESSLER6)		
Negative	81,437 (80.5%)	–
Positive	19,710 (19.5%)	–
Loneliness (ULS3)		
Negative	52,742 (51.5%)	–
Positive	49,625 (48.5%)	–
Psychological Well-Being (DIENER - 8-56, 56 being the most positive)	–	44.8

3.3 Data Analysis and Visualization

Exploratory data analysis (EDA) was performed, comparing all relevant sleep variables with demographics, diagnoses, and other outcomes. The first set of tests was performed with demographics. Comparing sleeping behaviors by sex (male and female) showed little to no differences, besides females taking slightly longer to fall asleep than males on average, which was around a difference of 7%. No significant trends were found between sleeping behaviors and year in school or race/ethnicity, although some notable findings were discovered. White people slept around 8% more than the mean, while Hispanic/Latino and American Indians slept around 4 to 5% less than the mean, and African Americans/Black people slept around 10% less than the mean.

There were slightly more significant findings between GPA and sleep. The average amount of time taken to fall asleep increased consistently as GPA decreased, primarily from A to C letter grades. While the difference in time to fall asleep is not extremely large between different groups, the consistency is notable. The average amount of sleep on weekday and weeknights decreased as GPA decreased as well, although only slightly, and students who reported an F GPA averaged around the same amount of weeknight sleep as B students. RBMI also did not play an important role in sleep results; however, obese people did, on average, take a longer time to fall asleep and experience a shorter amount of sleep, though not by considerable amounts.

The relationship between average weeknight sleep and GPA is worth discussing as well. There is a trend present where students with higher GPAs tend to get more sleep, and the correlation itself is consistent all the way down to the C- range, with the exception of A and A+ students being pretty much equal in average sleep duration. The difference between A+ and C- GPA students, in this relation, is around 35 minutes of sleep per night, on average. This is just one of many factors that will be explored in this study in terms of what contributes to and results from various sleep qualities. Figure 2 displays this trend.

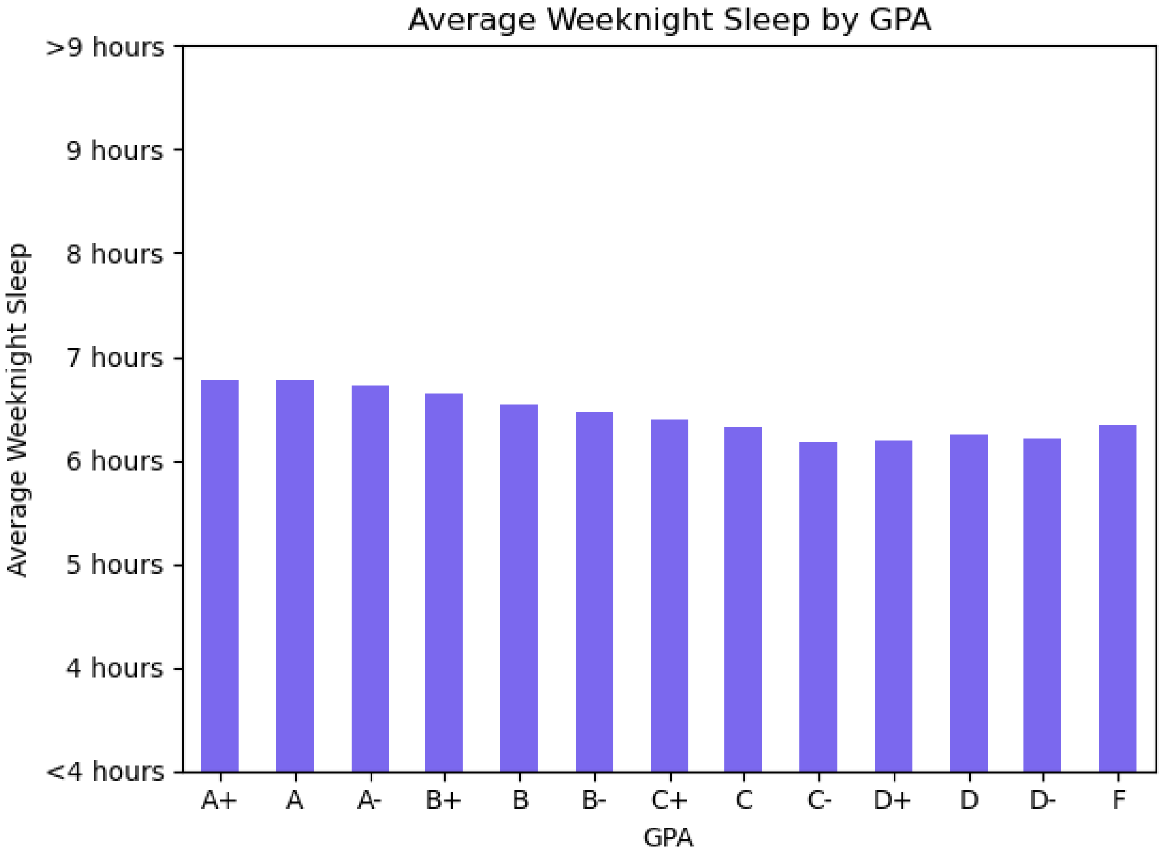


Figure 2: The relationship between GPA and average length of weeknight sleep.

Comparing the sleep latency and duration of students who live in fraternities and sororities with those who do not showed mostly negligible correlation. On weekend nights, the students living in fraternities and sororities do get slightly less sleep, but only by somewhere around 10 minutes less on average, which is nominal.

In terms of the connection between the presence of negative mental conditions and sleep, there are some interesting findings. To begin, the spring 2024 ACHA-NCHA data shows that out of roughly 100,000 American college students, 13.8% have been diagnosed with ADD/ADHD, 1.54% with an alcohol/drug use disorder, 35.2% with anxiety, 26.74% with depression, 7.17% with insomnia, and 2.36% sleep apnea. Figure 3 provides a visualization of these numbers.

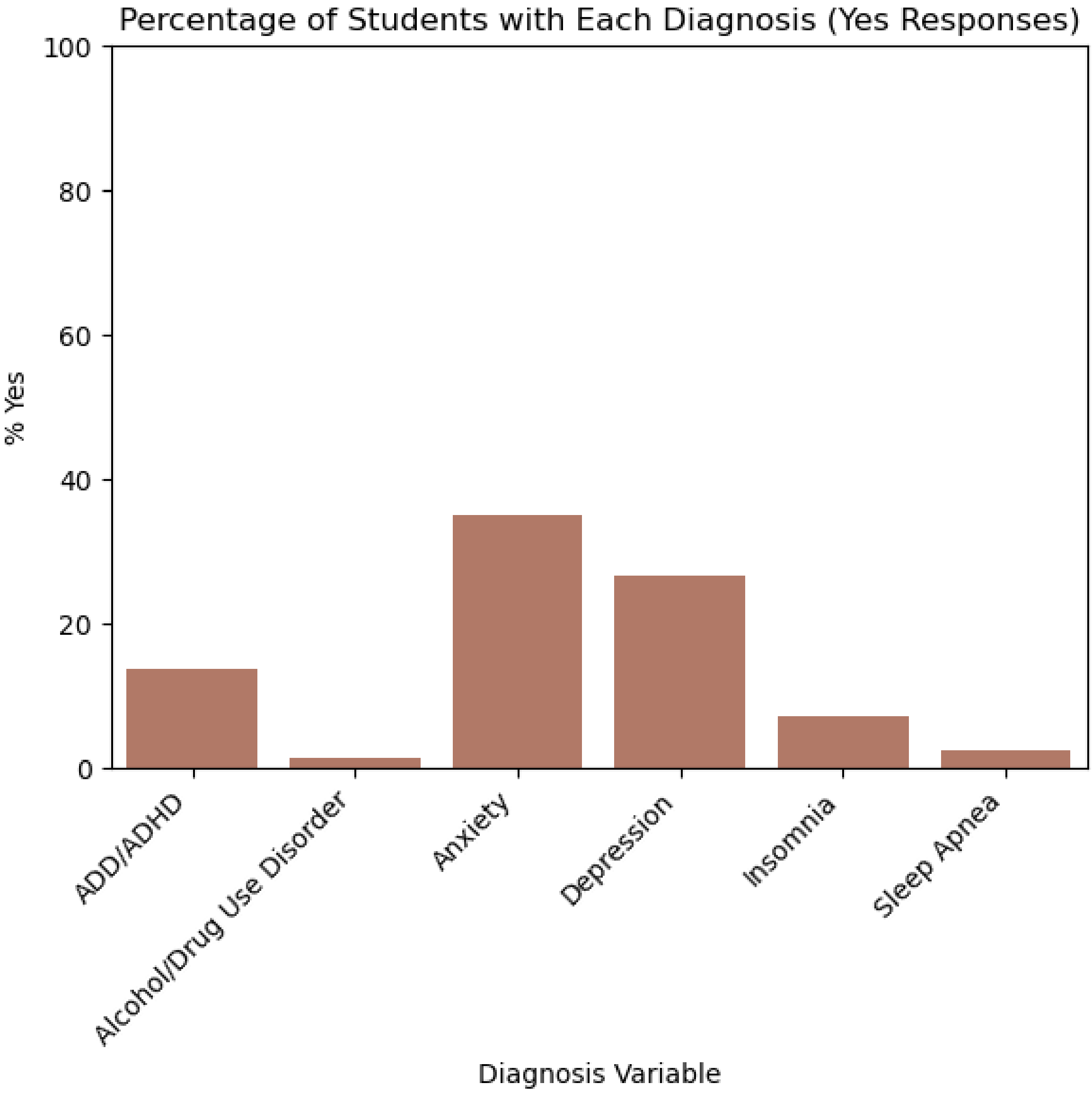


Figure 3: The percent of students who have been diagnosed with certain conditions.

People who have been diagnosed with ADD/ADHD, alcohol or drug abuse/addiction, anxiety, depression, insomnia, and sleep apnea all showed greater difficulty falling asleep than those without a negative mental condition. This was reflected most strongly in students with depression and insomnia, and a lack of sleep was reflected most strongly in students with insomnia and sleep apnea, as expected. Among the effects of sleeping difficulties on academics, 51% of students reported experiencing no sleeping difficulties, 26% reported experiencing sleeping difficulties but did not experience any undesired academic side effects, 22% of students experienced sleep difficulties and reported negative academic effects, with roughly 90% of this group (20% of total respondents) stating that their sleep difficulties contributed to poor performance in their classes, and the other 10% of this group (2% of total respondents) stated that their sleep difficulties caused their degree progress to be delayed.

There is a clear negative relationship between students taking longer to fall asleep and those who rate their overall health lower. For students who rate their overall health as ”excellent”, 53.7% say they typically fall asleep in under 15 minutes, while only 19.5% take over 30 minutes, whereas for students who rate their overall health as poor, only 21.8% fall asleep quickly, and 58.2% take over 30 minutes on average. Figure 4 shows this negative relationship quite clearly.

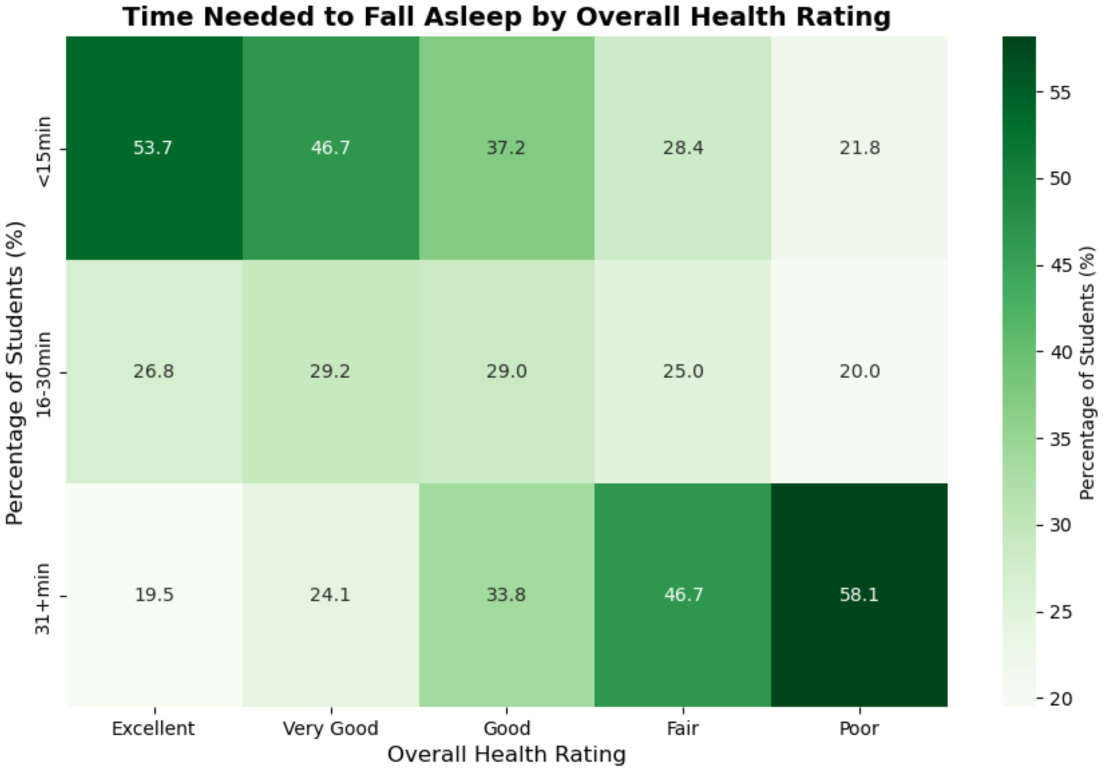
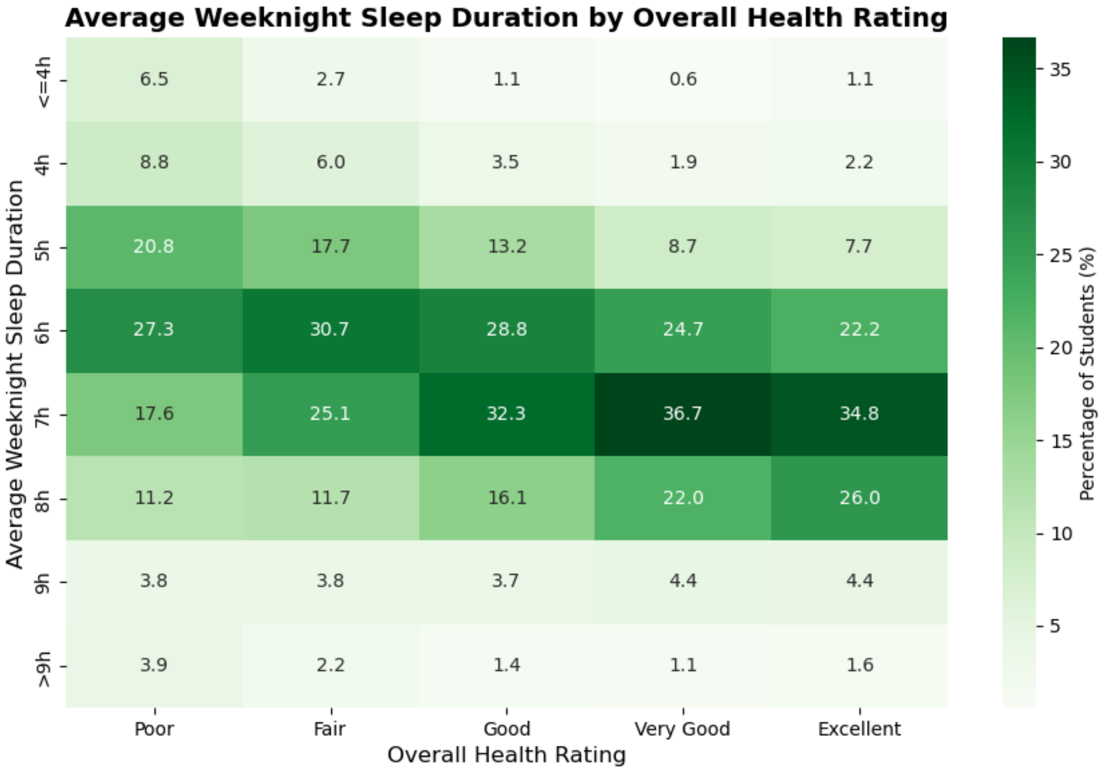


Figure 4: The average time needed to fall asleep across different self-reported overall health scores.

Students with worse overall health also got fewer hours of sleep on both weeknights and weekend nights than those with better overall health, with weeknight sleep having a stronger connection. It is worth noting, however, that these values accounted only for the students’ last two weeks of sleep, and were estimated and self-reported by students rather than being scientifically tracked. Figures 5 and 6 visually demonstrate this relationship, one through the medium of a heatmap and the other through the medium of a stacked bar chart. The relationships at play here suggest the idea that around 7-8 hours of sleep per night may be the optimal amount of sleep for the average college student. Approximately 7 hours of sleep is shown to be the most reported sleep duration for students who answered ”excellent”, ”very good”, or ”good” in terms of their overall health, with 8 hours also showing strong prominence among those with satisfactory overall health ratings.



	Poor	Fair	Good	Very Good	Excellent
<=4h	6.5	2.7	1.1	0.6	1.1
4h	8.8	6.0	3.5	1.9	2.2
5h	20.8	17.7	13.2	8.7	7.7
6h	27.3	30.7	28.8	24.7	22.2
7h	17.6	25.1	32.3	36.7	34.8
8h	11.2	11.7	16.1	22.0	26.0
9h	3.8	3.8	3.7	4.4	4.4
>9h	3.9	2.2	1.4	1.1	1.6

Figure 5: The average amount of weeknight sleep across different self-reported overall health scores (Heatmap).

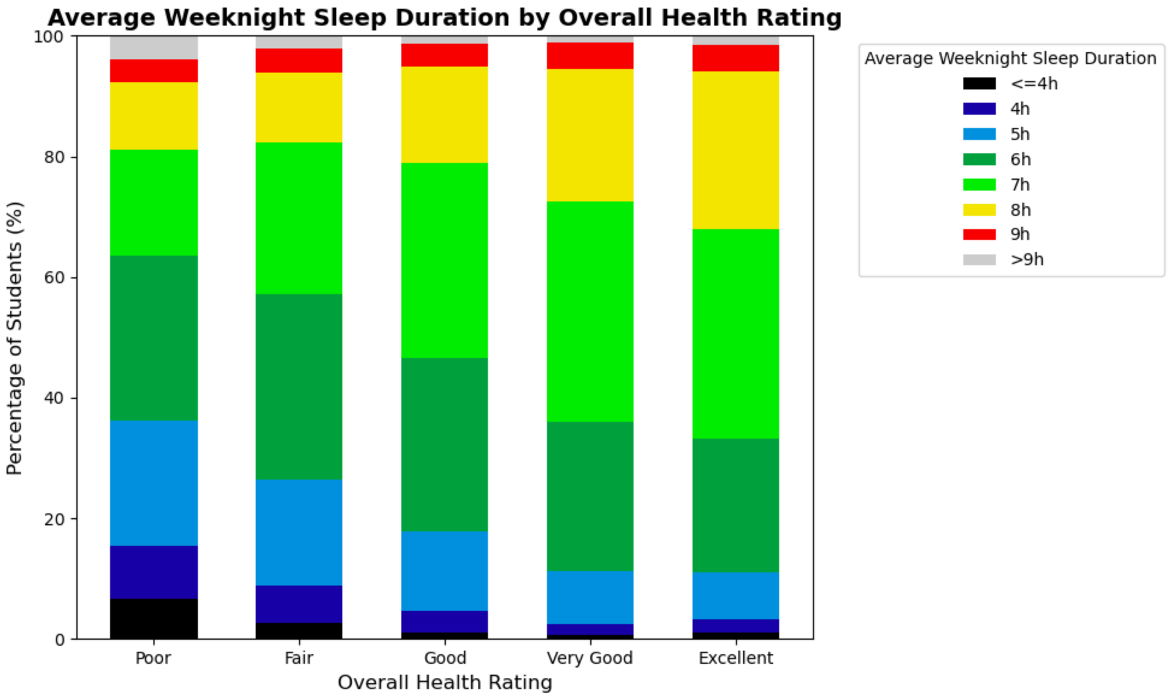


Figure 6: The average amount of weeknight sleep across different self-reported overall health scores (Stacked Bar Chart).

The amount of time that students spend doing physical activity did not have a very strong correlation with time taken to fall asleep, nor did it with weekday and weekend night sleep duration. Some minor correlations here, however, include moderate activity being slightly positively associated with optimal weeknight sleep of around 7-8 hours and slightly negatively associated with optimal weekend night sleep, as mean weekend night sleep slightly decreases as physical activity rises.

When comparing sleep with partying, however, the correlations are stronger. In terms of weeknight sleep, the mean duration of sleep decreases as partying increases. Students who tend to party for longer periods of time are more likely to get less than 5 hours of sleep on weeknights. On weekends, mean sleep also decreases as partying increases, but more considerably. Students who spend less time partying spend more time getting sleep that falls within the optimal range, according to the literature. Based on the data, the time taken to fall asleep also increases slightly with more time spent partying. Figure 7 shows the correlations between the average duration of sleep during the weekend and the amount of time spent partying per week.

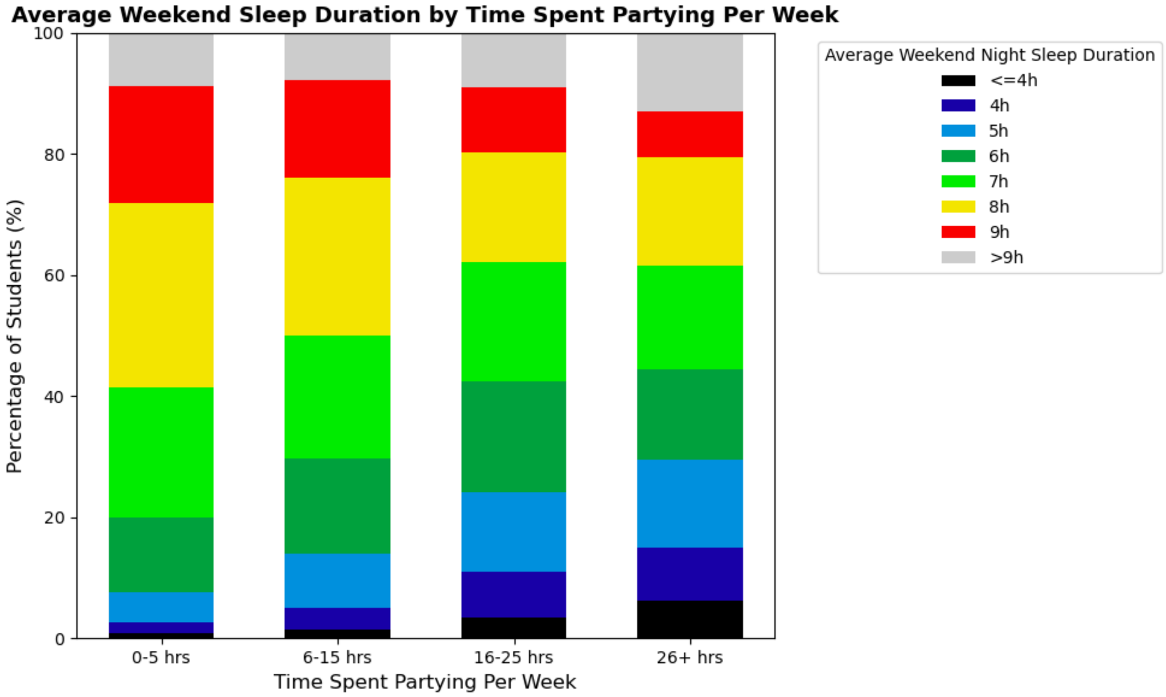


Figure 7: The average amount of weekend sleep compared to number of hours spent partying per week.

Another section of questions within the survey dealt with traumatic experiences that happened in the lives of students over the last 12 months. From these questions, correlations were investigated between the occurrences of sexual assault, rape, and stalking and the sleep variables that were being used in each of these analyses. Correlations were notable among each of these groups, especially those who were victims of rape. For students who had experienced sexual assault, falling asleep took roughly 5 minutes longer on average, and sleep duration was shorter on both weekends and weeknights by around 15 minutes on average, compared to the students who had not experienced sexual assault in the last 12 months. For students who had been a victim of stalking in the last 12 months, similarly to those who had been sexually assaulted, the time taken to fall asleep was about 5 minutes longer on average, and these students' sleep duration was around 25 minutes shorter on average, compared to those who had not experienced stalking in the last 12 months. These patterns are relatively mild but certainly notable. Most notably, however, is that for students who experienced rape in the last 12 months, the time taken to fall asleep was also around 5 minutes, but the duration of sleep on weeknights and weekends was, on average, about 30 minutes less than those who had not experienced rape in the last 12 months. Between those without rape trauma and those with rape trauma, the difference in sleep time on average is 6 hours 41 minutes compared to 6 hours 13 minutes on weekdays, respectively, and 7 hours 37 minutes compared to 7 hours 5 minutes on weekends, respectively. Figure 8 displays the relationship between students who were victims of rape in the last 12 months and their weekend night sleep duration. There is little to no difference between weeknight and weekend night sleep duration in each of these relationships investigated; however, typically the stronger relationship is being chosen for visualizations, which is weekend nights in this case.

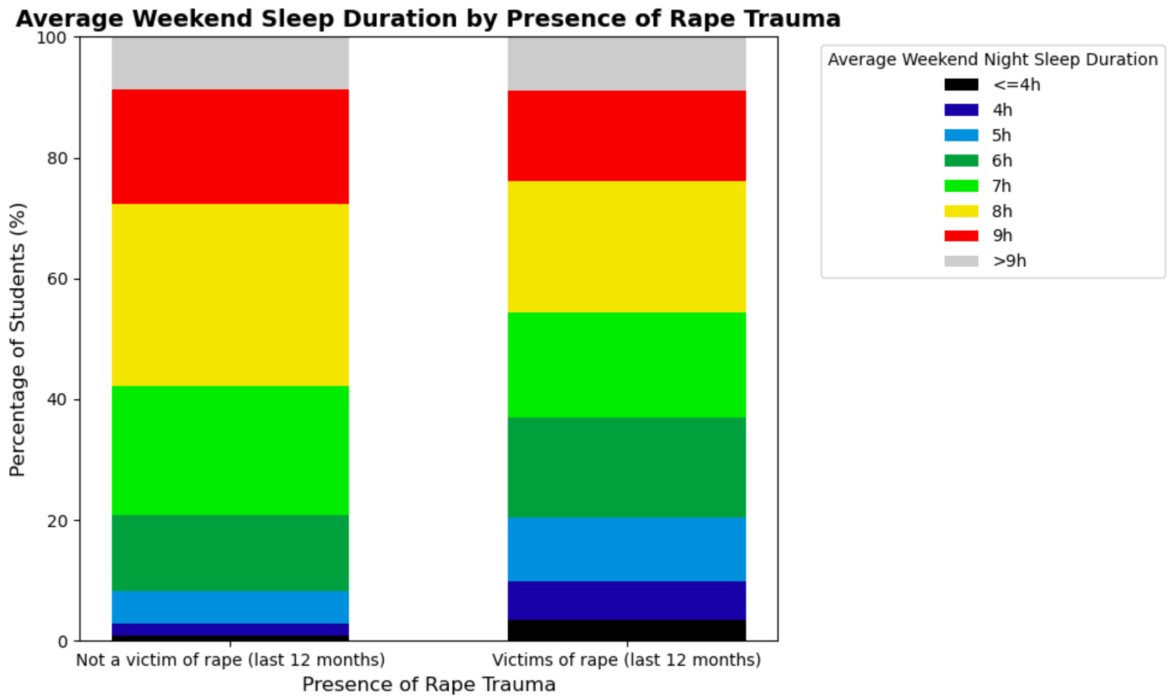


Figure 8: The difference in weekend sleep duration between those without rape trauma and those with rape trauma.

Students who took the survey were asked a question about whether they felt engaged and interested in their daily activities. Those students who said that they felt engaged and interested in their daily activities fell asleep faster on average, and also had more optimal sleep durations throughout the week, indicating that students who feel more interested and engaged in their daily lives have healthier sleep patterns. Those students who feel detached take longer to fall asleep, and are around 17% less likely (55% versus 38%) to get optimal sleep on weekdays and around 10% less likely (53% versus 43%) to get optimal sleep on weekends.

Another question that students answered in the survey was about their resilience, and ability to bounce back after illness, injury, and hardship. Among students with greater resilience there is a correlation with falling asleep quicker as well as getting more optimal sleep. The proportion of students who get 7-8 hours of sleep per night rises from around 38% to 55% on weekdays and 39% to 53% on weekend nights as resilience increases, with the group of people who sleep under 5 hours per night reducing from around 29% to 15% on weekdays and 20% to 8% as resilience rises. Also, the average score given by students who answered often resilient and very often resilient was a 3.0 on weekend nights, which is exactly the rating associated with 7-8 hours of sleep, which continues to be the most common answer among the most optimal health answers. Figures 9 and 10 both display a visual of the weeknight sleep duration compared to resilience scores, with Figure 9 being a heatmap and Figure 10 being a stacked bar chart. Also note that up to this point so far, results for students averaging 9 or more hours of sleep are mixed. There have not been any notably strong correlations thus far in this research, however the data here reveals that students certainly do use the weekends to catch up on sleep, as there are around 4-5 times more students getting 9 or more hours of sleep on the weekend nights as there are on weekday nights in the majority of these analyses (roughly 25-30% versus 5-8%, respectively).

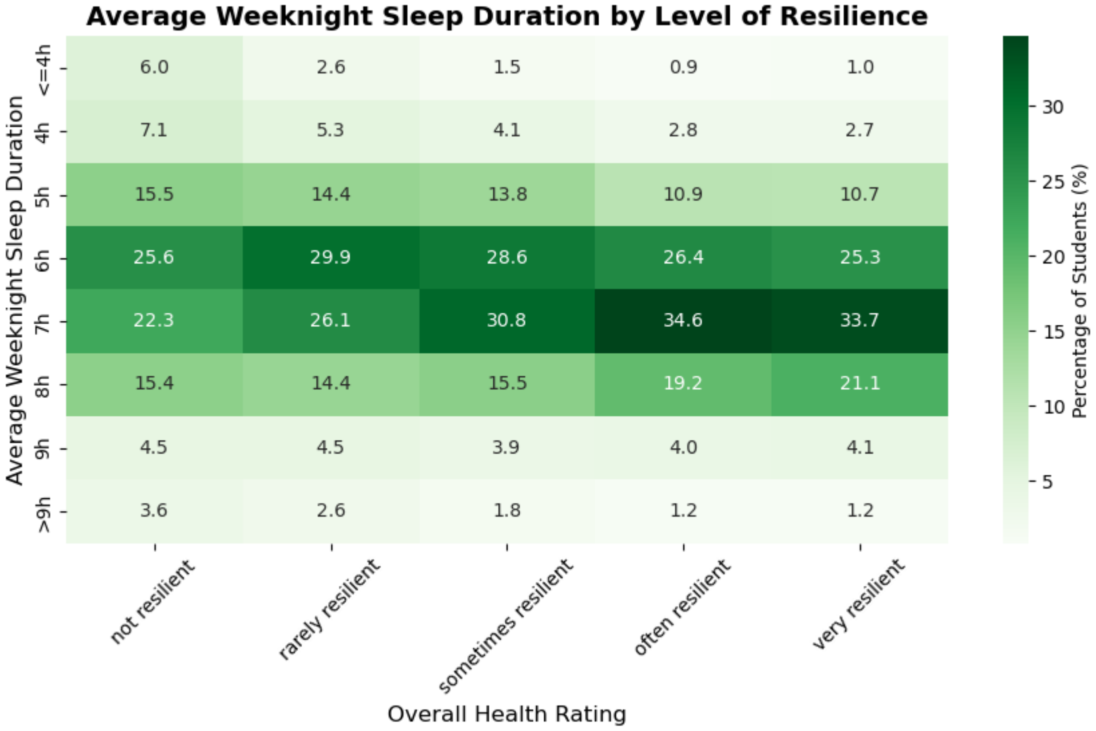


Figure 9: The average amount of weeknight sleep compared to self-reported resilience ratings (Heatmap).

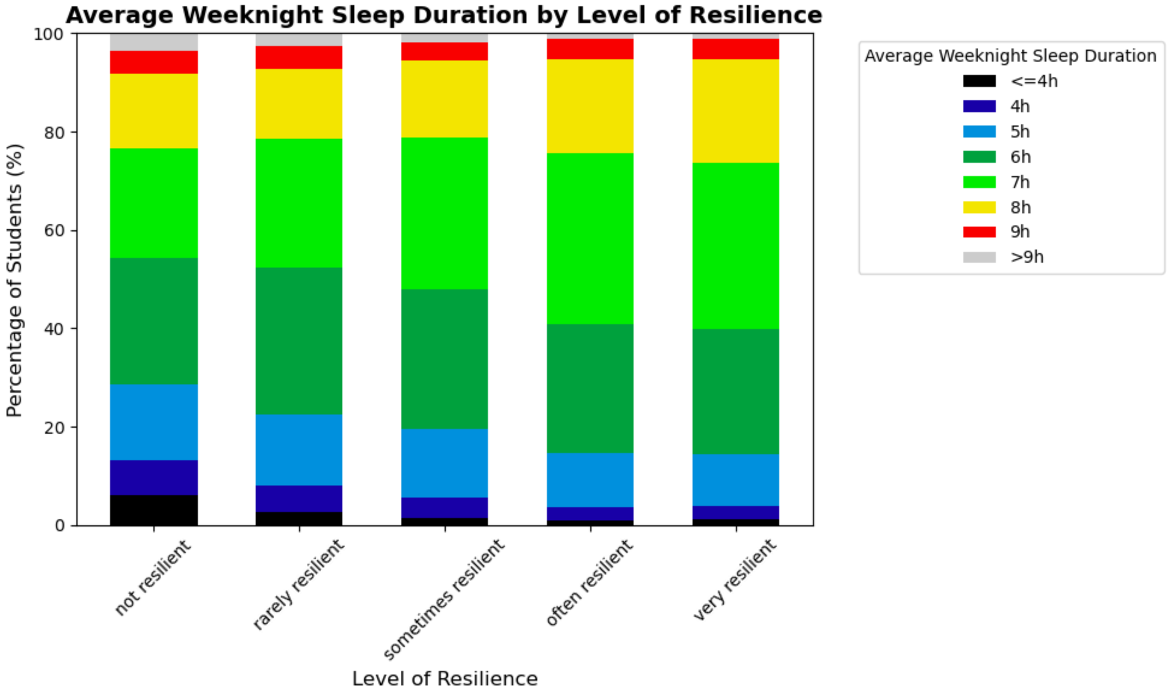


Figure 10: The average amount of weeknight sleep compared to self-reported resilience ratings (Stacked Bar Chart).

The correlations revealed between stress and sleep were strong. As stress increases, falling asleep becomes more difficult, with high-stress students being over twice as likely to take more than 30 minutes to fall asleep as compared to low-stress students (43% versus 20% respectively). In terms of the relationship between weeknight sleep duration and stress, sleep duration peaks around low-moderate stress and sharply declines at high stress levels. The size of the group with 7-8 hours of sleep drops from 63.7% at low stress to 53.7% at moderate stress, and then to 38.7% at high stress. The group with less than 5 hours of sleep more than triples between low and high stress, increasing from 8.4% to 26.5%. This indicates that having high stress often contributes to significantly shorter sleep on weeknights. Interestingly, the group with no stress did not perform the best, objectively. The no-stress group had more members of the ”less than 5 hours” group and less members in the ”7-8 hours” group than the low-stress and moderate-stress groups. The group above 9 hours represents a small fraction of the population and had narrowing representations as stress increased, moving semi-consistently and starting at 9.7% with no stress and ending at 4.4% with high-stress. On weekend nights, many of the same patterns repeated themselves; however, the high-stress group does not show nearly as sharp of a decline, and even has fewer members in the ”less than 5 hours” group and slightly more members in the ”7-8 hours” group. This further seems to indicate that students often use the weekend to rest up and recover, especially with the proportions of the ”9 hours and above” group seeing around 4-5 times the size on weekend nights. These insights also seem to indicate that having no stress might actually be sub-optimal for sleep, as the groups with low and moderate stress saw more optimal sleep durations, although having similar durations for time taken to fall asleep. It is also worth mentioning the breakdown of specifically the stress variable, where 1,863 students (1.8%) reported having no stress, 22,371 students reported having low stress (21.8%), 51,407 students reported having moderate stress (50.1%), and 26,938 students reported having high stress (26.3%). Figure 11 illustrates the sharp decrease of sleep duration on weeknights for high-stress individuals. Also note that this visualization style was chosen for many of these correlations as it is able to measure the relationships between these variables while also accounting for the proportions of students per response. The consistent color coding was chosen because of the consistent use of weeknight and weekday sleep.

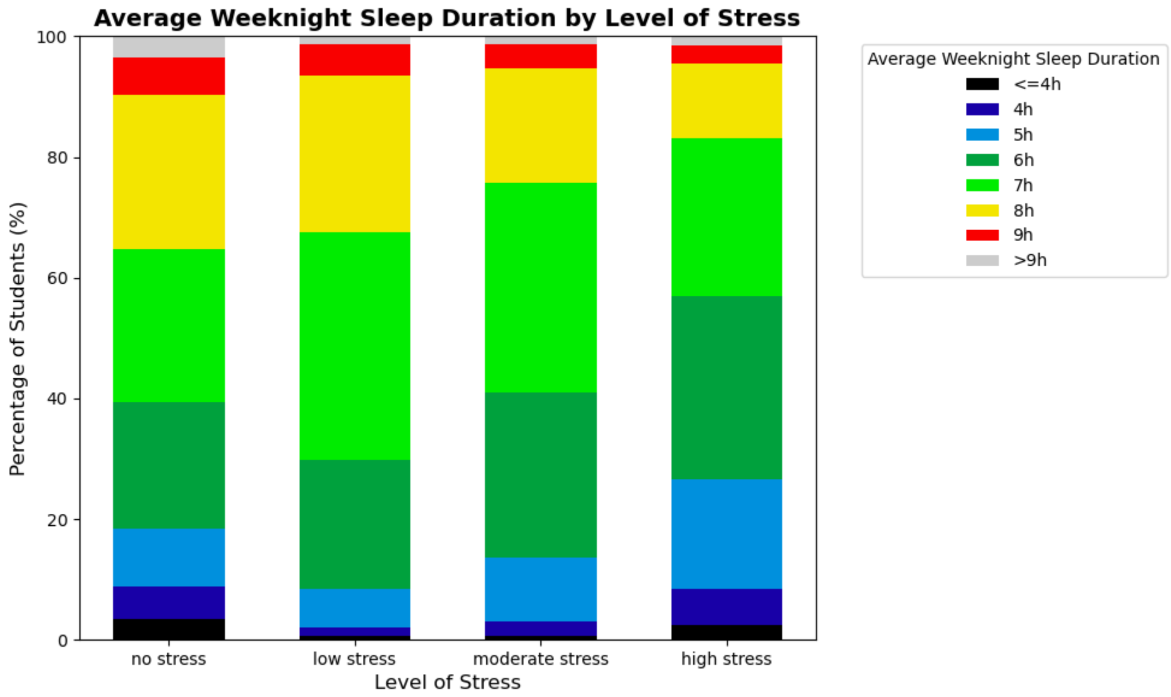


Figure 11: The average amount of weeknight sleep compared to self-reported stress ratings.

The psychological distress levels of students were measured using the Kessler Distress Scale, a validated measure of serious psychological distress (SPD). For students who scored having SPD, it took longer to fall asleep on average. There were nearly double as many students who took over 30 minutes to fall asleep in the group with SPD, compared to those without SPD. On weekdays, 58.1% of those who scored positive for SPD were in the range of getting 6 or fewer hours of sleep per night compared to just 38.7% for those negative for SPD. Comparably, only 36.7% of individuals with SPD reported obtaining 7-8 hours of sleep, whereas for individuals without SPD, the size of this group was 55.7%. For weekend nights, similar to the majority of previous relationships discussed, the correlations remain but are weakened. Interestingly, yet again, the size of the group who obtains 9 or more hours of sleep per night on average is roughly equal for both those negative and positive for SPD, but is multiplied in size around 4.5 times on weekend nights. This provides more support to the idea that many college students are stressed and perhaps overworked during the week and practice compensatory sleep behaviors on weekends.

The UCLA loneliness scale was also utilized in this study to measure how disconnected, lonely, and isolated students felt, placing them into two groups: one for those negative for loneliness and the other for those positive for loneliness, similar to what was done with the Kessler Distress Scale discussed previously. The connections between sleep and loneliness are notable, although not as strong as some previous correlations. Students who feel lonely do tend to take longer to fall asleep and also get less sleep on both weeknights and weekend nights than those who were not classified as lonely. On weekdays, the difference between those in the 7-8 hour range is around 10%, with 46.9% of those who classify as lonely getting 7-8 hours of sleep per weeknight, compared to 56.6% from the non-lonely category.

Another mental health variable used in this study is the Subjective Well-Being Scale designed by Ed Diener in 2009, which gives students a score of 8 to 56 based on a handful of questions that report a person's psychological well-being, with higher scores representing higher quality levels of well-being and overall life satisfaction. The findings here reinforce what has been previously discussed, with those who scored higher having less trouble falling asleep and also getting sleep closer to the 7-8 hour range. Figure 12 displays this relationship with weekday sleep duration. Note that for just this case, a score of 2.0 on the y-axis indicates 6 hours of sleep on average, where a 3.0 indicates 7-8 hours of sleep on average, and a 1.0 indicates 5 or fewer hours of sleep on average.

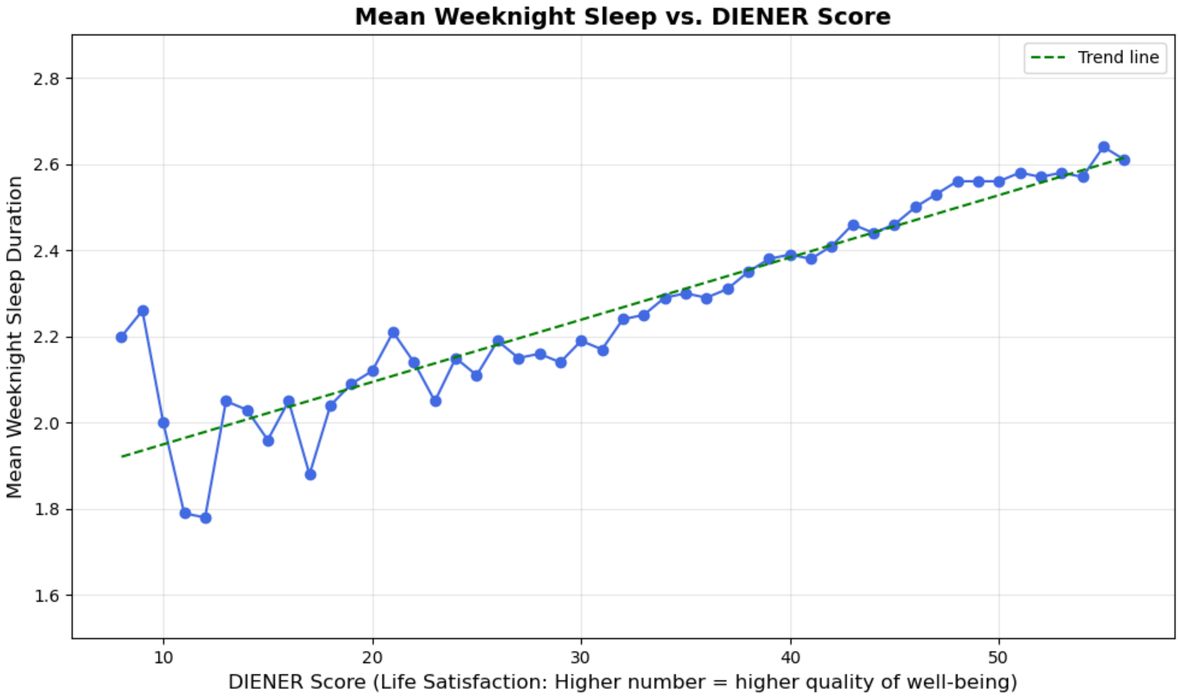


Figure 12: The average amount of weeknight sleep compared to DIENER scores (indicates quality of well-being).

There are five additional questions included in the correlation analysis that were not mentioned earlier. Each begins with "On how many of the last 7 days did you..." The questions are:

1. "Wake up too early in the morning and couldn't get back to sleep?"

2. "Feel tired or sleep during the day?"

3. "Have an extremely hard time falling asleep?"

4. "Get enough sleep so that you felt rested?"

5. "Take a nap?"

For this set of sleep variables, there were no significant correlations with any of the other variables. There was, however, a correlation between each of these questions and people stating that they have experienced sleep problems, which is logical, with having an extremely hard time falling asleep and not getting enough sleep to feel rested as the most relevant. There was also a slight correlation where students who party more often tend to have a more difficult time falling asleep.

Correlations were also investigated between students with sleep apnea or insomnia and all of the variables discussed above. The sample size of students with sleep apnea in this study is 2,394, which is 2.36% of the total, and the sample size of students with reported insomnia in this study is 7,239, which is 7.17% of the total.

For students with sleep apnea, the most notable correlation was with substance abuse, as students with a substance abuse diagnosis were 4.5 times more likely to report sleep apnea than students without a substance abuse diagnosis (12.1% versus 2.2%). Students with depression were around 4 times more likely to report sleep apnea than those without depression (5.34% versus 1.26%). Students with anxiety were around 3.5 times more likely to report sleep apnea than those without anxiety (4.38% versus 1.26%). For students reporting negative academic impacts, the risk was almost twice as high, and for students who reported ADD/ADHD, the risk was almost three times as high. Figure 13 shows the rates of comorbidity among students with sleep apnea. Note that the scale of the Y-axis is equal on both charts, ranging from 0% to 35%.

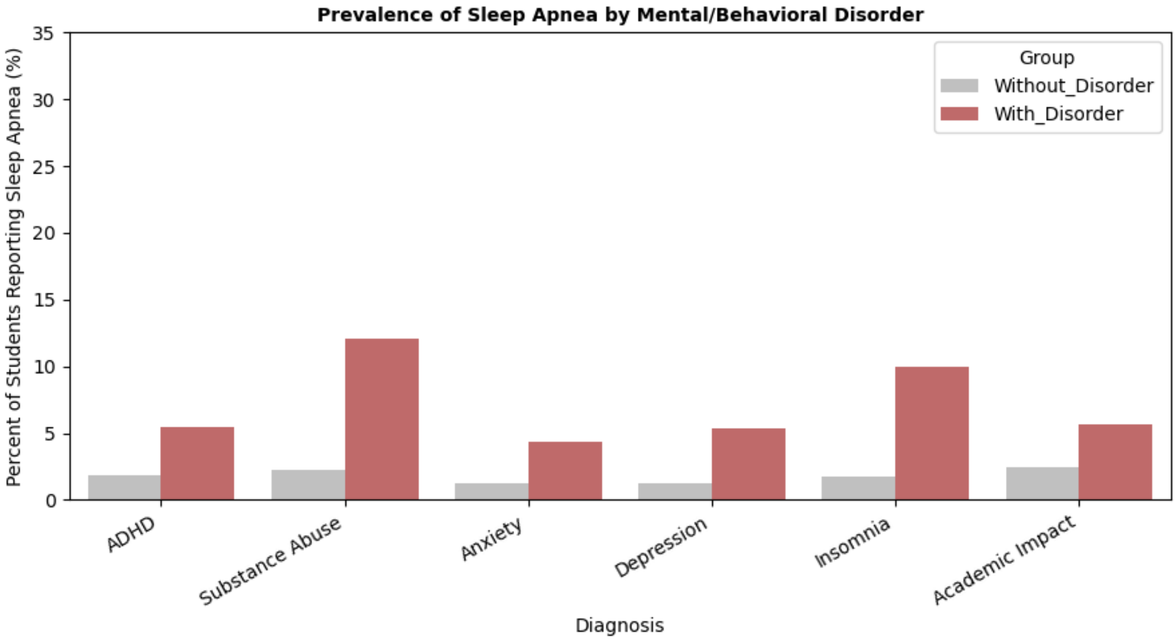


Figure 13: The percentage of students without each diagnosis who reported sleep apnea, compared to the percentage of students with that diagnosis who reported sleep apnea.

Insomnia had many stronger correlations across the board. The strongest correlations here were between anxiety, depression, and substance abuse. Students with anxiety were around 9 times more likely to report insomnia than students without anxiety (16.95% versus 1.87%). For depression, the ratio was very similar, with students with depression being around 8.5 times more likely to report insomnia than those without depression (20.34% versus 2.37%). Furthermore, students who were diagnosed with a substance use disorder were nearly 5 times more likely to experience insomnia compared to those without a substance use disorder (33.33% versus 6.75%). For students with sleep apnea, ADD/ADHD, and those who experienced negative academic impact from their disorders, the chances of reporting insomnia were around 4 times more likely than students without these disorders. All of this data indicates that students with insomnia often have other mental disorders. As indicated by the review of the literature, as well as the findings in this study, the relationship between sleeping habits and sleep disorders is bidirectional. For example, a mental disorder like depression may contribute to insomnia symptoms, while depressive symptoms may also arise as a result of the insomnia itself, potentially caused independently from the depressive symptoms. Figure 14 shows the rates of comorbidity among students with insomnia. Again, the scale of the Y-axis is equal on both charts (insomnia and sleep apnea), with that range being 0% to 35%.

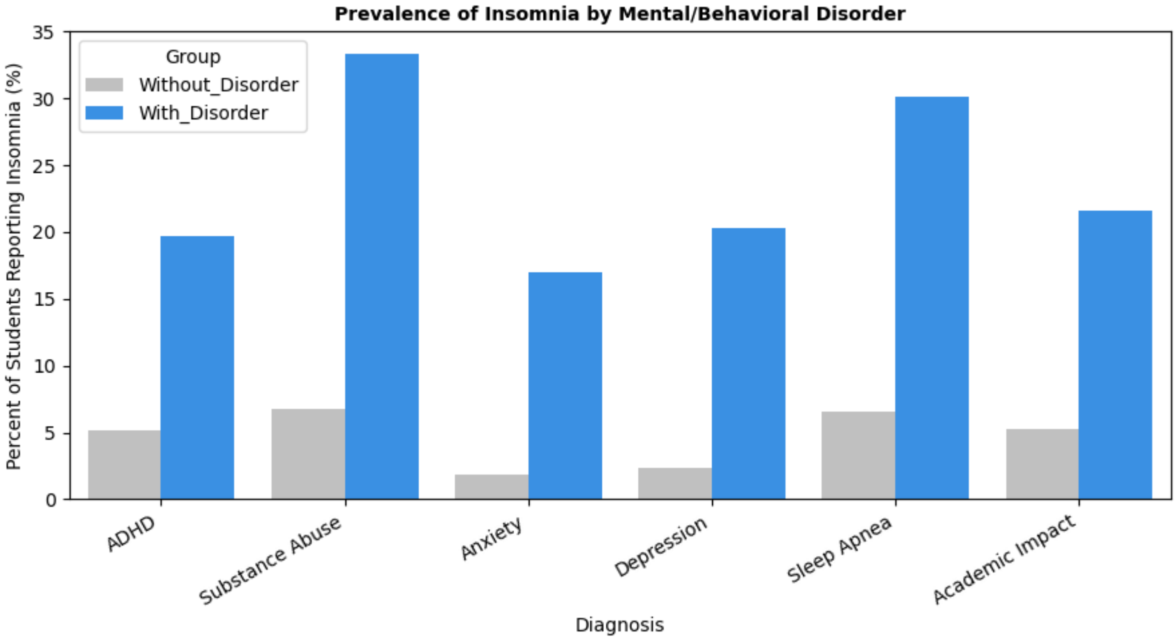


Figure 14: The percentage of students without each diagnosis who reported insomnia, compared to the percentage of students with that diagnosis who reported insomnia.

Interestingly, students with diagnosed sleep apnea are around 3.5 times more likely to report having insomnia than those without sleep apnea (30.1% versus 6.6%), but students with insomnia are around 4.5 times more likely to have sleep apnea than those without insomnia (9.93% versus 1.78%). This pattern suggests that while these conditions often co-occur, insomnia may be a stronger indicator of overlapping sleep issues than sleep apnea is. Correlation analyses reveal that insomnia demonstrates stronger associations with other mental disorders than sleep apnea. Nonetheless, both sleep conditions exhibit markedly higher comorbidity with mental health diagnoses relative to students without a sleep disorder, which highlights their shared psychological burden. Next, the building and testing of a predictive model will be discussed. Refer to Appendix C: Analyzing for the Python code used for this section of the pipeline.

3.4 Model Building and Testing

3.4.1 Linear Regression

Now that data has been collected from previous literature and the firsthand analysis that has been conducted in this study, the next step is developing a model. A predictive model was chosen to estimate how much sleep a college student gets based on many of the variables discussed earlier. Linear regression was chosen for this as it is a powerful yet simple method for modeling the relationship between one independent variable (weeknight sleep duration) and a number of dependent variables. The relationships are modeled using linear predictor functions that use model parameters that are estimated from the data [26]. This approach has been used in previous sleep research frameworks, such as in the stress-sleep theory, which proposes a nearly linear relationship between stress and sleep duration [27], the bidirectional model of sleep and mental health, which suggests consistent associations between anxiety, depression, and substance use disorders and sleep duration, and cognitive-behavioral therapy for insomnia (CBT-I), which uses psychological symptoms to predict sleep outcomes [28]. The goal here is to determine how these predictors and changes in them are associated with changes in sleep duration. Adjusted Coefficient of Determination (Adjusted R^2) and Mean Absolute Error (MAE) values of the model were used to provide insights into model performance. The adjusted R^2 value indicates how much of the variance in the target variable can be explained by the predictors in the model, thus also indicating how much of the results come from factors outside of what is being tested. The adjusted R^2 value ranges from 0 to 1, where 0 means that the model explains none of the variation and 1 means that the model is able to explain all of the variation. Adjusted R^2 was chosen because normal R^2 will always increase or stay the same when more predictors are added, which means it fails to fully explain if newly implemented predictors are failing to improve the model's explanatory power, whereas with adjusted R^2, the model will be penalized when newly implemented predictors fail to improve the model's performance, genuinely showing if new predictive information is being contributed. Equation 1 below shows how to calculate adjusted R^2.

$$\text{Adjusted } R^2 = 1 - (1 - r2) * ((n - 1) / (n - p - 1)) \tag{1}$$

where R^2 is the sample R-squared, n is the total sample size, and p is the number of independent variables.

MAE is the average magnitude of errors between the predicted and actual sleep durations, where a lower MAE means better predictive accuracy. In this dataset, weeknight sleep duration was measured using a categorical coding system in which each numerical code corresponds to an hour of sleep (the code "3" represents 5 hours, "4" represents 6 hours, and so on). Therefore, an MAE of 0.50 would mean that, on average, the model's predictions differ from the actual reported sleep duration by 30 minutes, since having a MAE score of 1 would represent an average difference of one hour. MAE was computed using the mean_absolute_error function included in the Python Sci-Kit learn module.

Apart from the Sci-Kit Learn module in Python, the Pandas module was also utilized in the process of creating this model. Similar to the analyzing section, variables were broken into groups. For the sake of modeling, there are 5 groups, where each group was added into the model one at a time, and the individual weights of each variable were returned, essentially rebuilding the model as each group is passed in. These 5 groups include:

- Psych - resilience, stress, well-being/life satisfaction (DIENER), serious psychological distress (RKESSLER6), and loneliness (RULS3)

- Behavior - overall self-rated health, time spent doing physical activity, time spent partying, and time taken to fall asleep

- Diagnoses - ADHD, substance abuse, anxiety, depression, and sleep apnea

- Trauma - victims of sexual assault, rape, and stalking

During the initial data analysis and visualization phase, rows with missing values were temporarily kept and handled on a case-by-case basis so that if one column was missing data, the rest of the row could still contribute to other correlations. For building the model, however, all missing values had to be addressed across all predictors and the target variable. This process resulted in a sample size of 88,408 by the end of the model's operating processes, as with each group being passed into the model, the rows that included missing values in any of the columns within that group were dropped, appropriate for regression fitting. So, in a sense, the "model" is a series of 4 linear regression models that build upon each other as predictor groups are added. The next section will address the results of testing the model.

The model was executed, and the raw results were gathered. The adjusted R^2 of the model improved with each new group of predictors added, indicating that each group furthers the effectiveness of the model. These levels of effect differed significantly in some cases, however. Starting the model with the psychological group already explained 5.5% of the variation in weeknight sleep duration, suggesting that students' resilience, stress, well-being, psychological distress, and loneliness contribute 5.5% of the total equation for what influences sleep duration. Coming after psychological factors was the behavioral group, which explained 2.3% of the variance in weeknight sleep duration. The trauma group, including victims of sexual assault, rape, and stalking, contributed 1.9% to what constitutes a student's weeknight sleep duration. The diagnoses group showed a contribution of only 0.004%. The likely reason this group contributed only slightly is likely due to many of the aspects of certain important disorders, such as depression and insomnia, being covered in the psychological and behavioral group. In the analyses of the individual predictors, however, it will become clear that several of the variables in these groups did have a large impact by themselves regardless.

MAE decreased from 0.952 to 0.907, indicating that the model's sleep predictions were off by about 55 minutes at the end, with only around 2 or 3 minutes being cut off with each group of contributors passed into the model, which is logical as there are many things that contribute to one's sleep duration outside of the realm of the data used in this study such as screen time, screen time before bed, caffeine usage, academic workload, diet and exercise timing, and sleeping environment. Table 3 shows a summary of model performance with each predictor group that was added.

Table 3: Summary of Linear Model Performance (Predictor Groups - with Diagnoses Passed in First)

Predictor Group	**Adjusted R^2**	**MAE**	**Interpretation**
Psychological	0.055	0.952	Baseline model, meaningful explanatory power.
+ Behavioral	0.078	0.918	Adding behavioral factors improved prediction substantially (0.023).
+ Diagnoses	0.078	0.920	Mental/physical diagnoses only added marginal improvement (0.0004).
+ Trauma	0.097	0.907	Trauma related experiences improved the accuracy notably (0.019).

The model was run again to test how passing in the diagnoses group first affected the spread in variance. Table 4 displays the results of the model again, but with passing the diagnoses predictor group into the model first. Adjusted R^2 and MAE varied from the first model as groups were added in a different order, but running the model with the predictors in either order resulted in the same adjusted R^2 and MAE. When the diagnoses group is run first, an extra 2% of variance is explained, although the psychological groups and behavioral groups both lose out on 1% of variance explanation each, while the trauma predictors explained the same level of variance. The demonstrates that the diagnoses predictors do in fact have overlaps with psychological and behavioral variables. The overall adjusted R^2 value for the model is 0.097, which is 9.7%, meaning that 9.7% of the variability in sleep duration is explained by the predictors used in the model. This indicates that, based on this data, around 90.3% of the variability in sleep duration are explained by factors not present in the model, likely due to aforementioned factors such as screen time, screen time before bed, caffeine use, academic workload, diet and exercise timing, and sleeping environment. Unobserved errors, measurement errors, and random chance may also contribute to the equation of why this number is so low.

Table 4: Summary of Linear Regression Model Performance (Predictor Groups - Diagnoses First)

Predictor Group	**Adjusted R^2**	**MAE**	**Interpretation**
Diagnoses	0.021	0.978	Baseline model, small but meaningful explanatory power.
+ Psychological	0.066	0.936	Adding psychological factors improved prediction substantially (0.045).
+ Behavioral	0.078	0.920	Behavioral factors added notable improvements (0.012).
+ Trauma	0.097	0.907	Trauma related experiences improved the accuracy notably (0.019).

Looking specifically at the predictors themselves, some influenced the model much more than others. Table 5 displays these values from strongest to weakest based on their impact on the model. Note that these coefficient values represent the first model framework discussed, but are nearly entirely identical to the values received on the second run through of the model.

Table 5: Summary of Linear Regression Model Performance (Individual Predictors)

Rank	Predictor Variable	Coefficient	Interpretation
1	N3Q65A28 - Insomnia diagnosis	-0.338	Students with insomnia sleep much less on weeknights
2	N3Q65A35 – Depression diagnosis	-0.271	Depression is strongly associated with reduced sleep.
3	N3Q48 – Stress	-0.209	Higher stress leads to less sleep.
4	N3Q65A7 – Anxiety diagnosis	+0.200	Slightly higher sleep for those reporting anxiety (could be due to confounding or coping differences).
5	N3Q13 – Time to fall asleep	-0.199	Longer time to fall asleep correlates with less total sleep.
6	N3Q20G - Stalking victimization	-0.199	Victims of stalking report less sleep
7	N3Q20F - Rape victimization	-0.132	Victims of rape report less sleep
8	N3Q65A15 - Substance abuse diagnosis	+0.100	Unexpected positive... implies people with substance abuse issues get more sleep
9	N3Q65A2 - ADHD diagnosis	-0.096	Students with ADHD get slightly less sleep
10	RKESSLER6 – Serious psychological distress	-0.087	More distress correlates with less sleep
11	N3Q1 - Self-reported overall health	-0.087	Poorer overall health correlates with less sleep
12	N3Q20D - Sexual assault victimization	-0.069	Victims of sexual assault report less sleep
13	N3Q42AB – Resilience	-0.047	Slightly lower resilience leads to less sleep
14	RULS3 - Loneliness	-0.033	More loneliness leads to slightly less sleep
15	N3Q3I - Time spent partying	+0.016	Unexpected slight positive... implies weeknight sleep increases with time spent partying
16	N3Q3E - Time spent doing physical activity	+0.015	Slightly more activity correlates with slightly more sleep
17	DIENER – Psychological well-being	+0.009	Very subtle link being positive well-being and more sleep

From these values, it is apparent that students who have insomnia and depression, high stress, and those who have been victims of stalking tend to get less sleep. This model, however, produced an unsatisfactory result, as the adjusted R^2 value was too small to consider the model effective. From here, the study pivoted to the creation of a logistic regression model, which is different from a linear regression model in that a linear regression model predicts a numeric outcome, whereas a logistic regression model predicts a categorical outcome. Refer to Appendix D: Linear Regression Modeling for the Python code used for this section of the pipeline.

3.4.2 Logistic Regression and Additional Models

Logistic regression is another machine learning model archetype that is focused primarily on classification rather than prediction, dissimilar to linear regression [29]. A similar approach can be seen in Spielman's 3P Model of Insomnia [30] and the Hyperarousal Model of Insomnia [31]. For this logistic regression model, the survey respondents were divided into two groups. The first group is the ”suboptimal” sleepers. These are people who answered the question asking about their weeknight sleep duration over the last two weeks with a response that indicated 6 or less hours of sleep per night. The other group are people who responded with 7-9 hours of sleep per night. These groups were chosen due to information from the review of the literature and the firsthand analysis conducted in this study, where, in the majority of cases, sleep of 7-9 hours per night aligns with optimal outcomes, whereas 6 or fewer hours tends to align with more undesirable outcomes as well as contributors. Unfortunately, due to the question format in the survey, there is a gray area between 6 and 7 hours of sleep, which is just shy of what is typically considered optimal, but is not necessarily sub-optimal either. Looking back at the findings from the analysis, combining the 6 and 7 hour blocks gives relatively consistent data across different outcomes and contributors, but leans towards the optimal side. Also, people who reported greater than 9 hours of weeknight sleep over the last two weeks were not included, as based on both the literature and the analysis performed in this study, it would be difficult to label these people as optimal or suboptimal, as getting over 9 hours of sleep is typically seen as recovery sleep, often from things like illness. Note that not-a-number (NaN) values were coded as the mean of their column in order for them to still be able to be utilized.

Certain metrics were coded into this model to give insights into the model's performance. These include the accuracy of the model (what proportion of students are correctly classified) and a ROC-AUC score, which is the ability of the model to distinguish between classes. A confusion matrix and a classification report were also generated to give further insights into the model. The model resulted in an accuracy score of 0.615, indicating that this model was able to correctly classify 61.5% of students. Of the students predicted to get 7-9 hours of sleep, 62% actually do, according to the precision score, which is acceptable. The ROC-AUC score was 0.637, indicating poor-to-moderate predictive power, where random predictive power would have been a score of 0.5. This means that the probability that the model correctly ranks a random "optimal sleeper" above a "suboptimal sleeper" is 63.7%. The precision score calculates how many positive cases were correctly predicted out of the group of actual positive cases, which was 0.805 for the optimal group, indicating that the model catches 80.5% of optimal sleepers. The recall (sensitivity) score calculates how many positive cases were correctly predicted out of the group of actual positive cases, which was 62.3%, meaning that for the students predicted to get 7-9 hours of sleep, 62.3% actually do. The specificity value indicates that the model correctly identifies around 59.2% of suboptimal sleepers, compared to the recall value of 62.3%. The F1-score for the optimal group was 0.7, and is the harmonic mean of the precision and recall variables, and this number indicates balanced performance. Table 6 shows a summary of this data, and located below Table 6, Figure 15 displays the ROC curve for the logistic regression model.

Table 6: Summary of Logistic Model Performance

Metric	**Meaning**	**Interpretation**
Accuracy: 0.615	The overall percent of correct predictions	The model correctly classifies 61.5% of students
ROC-AUC: 0.637	Probability of ”optimal sleepers” being ranked over ”suboptimal sleepers”	AUC = 0.637 indicates poor-to-moderate predictive power (random is 0.5)
Precision (class 1 = optimal sleep)	The model captures 80.5% of optimal sleepers	Good, the model correctly captures the majority of optimal sleepers
Recall (class 1)	For students predicted to get 7-9 hours of sleep, 62.3% do.	Acceptable value
Specificity	For students predicted to get 6 or less hours of sleep, 59.2% do.	Acceptable value
F1-score (class 1)	The harmonic mean of precision and recall	0.70 indicates balanced performance

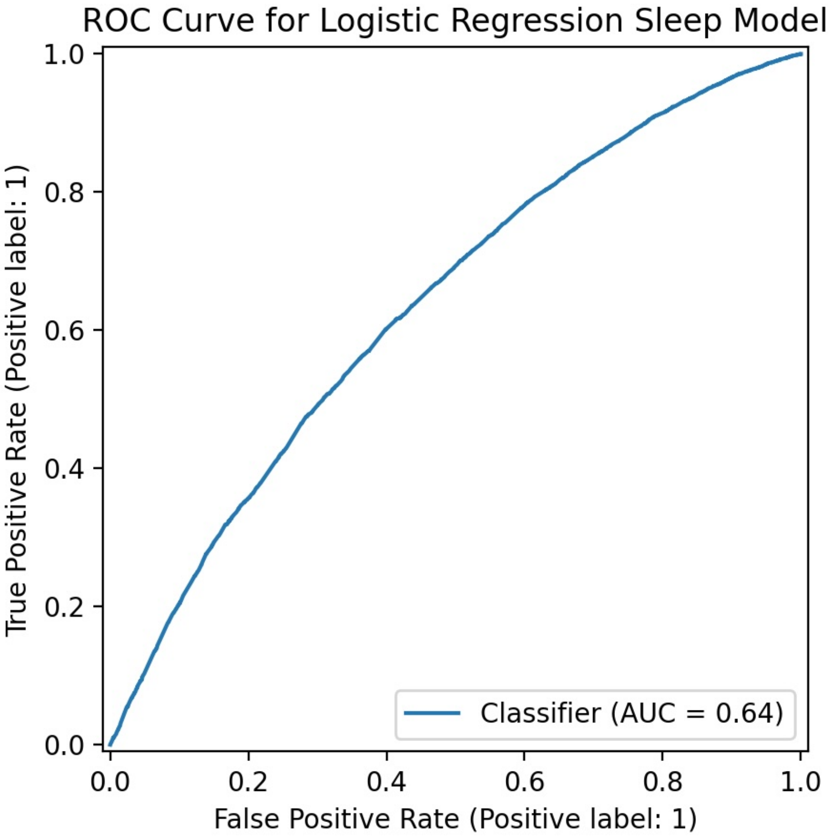


Figure 15: The ROC Curve of the Logistic Regression Model.

These aforementioned values are calculated using the confusion matrix resulting from the application of the regression model in Figure 16. From this matrix, there were 3,275 true negatives (TN), 5,628 false negatives (FN), 9,278 true positives (TP), and 2,254 false positives (FP). Equations 2-6 provide the equations for the accuracy, precision, recall, F1-score, and specificity metrics, and below Equations 2-6, Figure 16 shows the confusion matrix of the logistic regression model.

Accuracy = (TP + TN)/(TP + TN + FP + FN) ———— (2)

Precision = TP/(TP + FP) ———— (3)

Recall = TP/(TP + FN) ———— (4)

F1-score = 2 x (precision x recall)/(precision + recall) ———— (5)

Specificity = TN/(TN + FP) ———— (6)

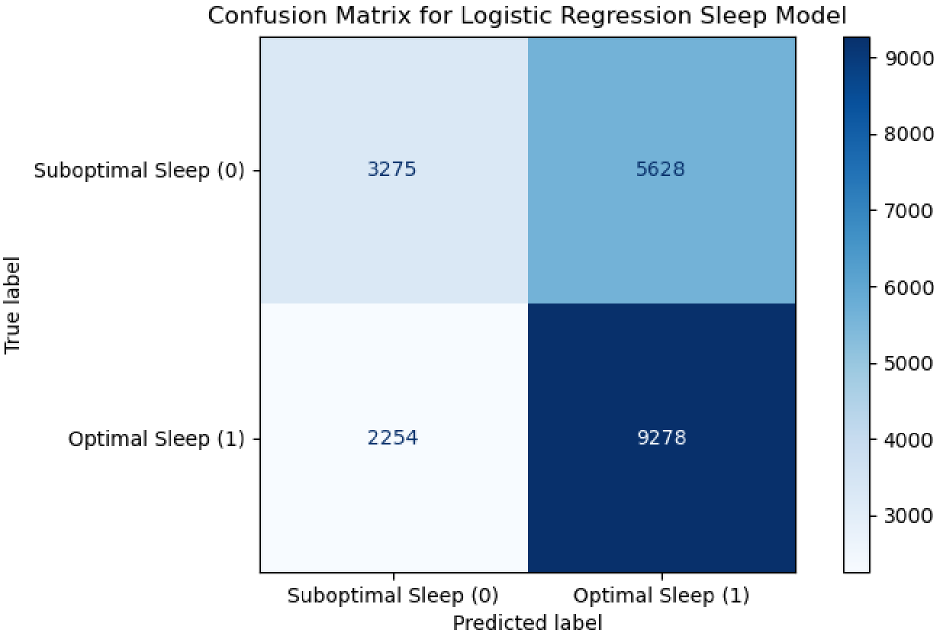


Figure 16: The confusion matrix of the logistic regression model.

In addition to the previous data, the coefficients and adjusted odds ratios were gathered from each of the individual variables to measure the impact that these variables had on the model.

The predictor with the strongest impact on sleep in the logistic regression model was stress level, where with each increase in stress level (no/low/moderate/high stress), the odds of optimal sleep decrease by 26%. Overall health rating was also a strong predictor, where with each improvement in self-reported overall health rating (poor/fair/good/very good/excellent), the odds of optimal sleep increase by 22%. Insomnia had a negative effect on optimal sleep, which was an expected relationship. In this case, the odds of optimal sleep were reduced by 15% when diagnosed insomnia was present. However, an unexpected relationship was the presence of diagnosed anxiety in students showing, 14% higher odds of obtaining optimal sleep. This is counterintuitive, especially since stress and anxiety are typically considered closely related, and anxiety had the opposite effect on predicting optimal sleep. Perhaps this is the result of coping strategies, treatment effects, or collinearity, such as anxiety co-occurring with increased awareness of one's stress. The existence of depression also increased the odds of optimal sleep, but less so at 3%. Student resilience scores (ability to bounce back from injury, illness, or hardship) showed a 4% higher odds with each level of resilience (not/rarely/sometimes/often/very often resilient), indicating that students who possess stronger coping and recovery skills tend to sleep slightly better. The existence of diagnoses of substance abuse, ADHD, and sleep apnea each contributed to lower odds of optimal sleep at 3%, 5%, and 6%, respectively. Victimization of sexual assault, rape, and stalking each contributed to lower odds of optimal sleep as well, at 2%, 3% and 7%, respectively. Lonely students have 8% lower odds of optimal sleep, indicating that social isolation negatively affects rest. A limitation of this model is that it only utilizes weeknight sleep duration and does not include weekend night sleep duration. These results are summarized concisely in Table 7 below.

Table 7: Logistic Regression Coefficients and Odds Ratios for Predicting Optimal Sleep

Variable (Survey Item)	**Coefficient**	**Odds Ratio**	**Brief Interpretation**
N3Q1 - (Overall health rating)	0.2026	1.2246	+22% odds of optimal sleep
N3Q65A7 - (Anxiety diagnosis)	0.1346	1.1440	+14% odds of optimal sleep
N3Q42B - (Resilience)	0.0418	1.0427	+4% odds of optimal sleep
N3Q65A15 - (Depression diagnosis)	0.0315	1.0320	+3% odds of optimal sleep
N3Q20D - (Sexual assault victims)	-0.0185	0.9817	-2% odds of optimal sleep
N3Q65A3 - (Substance use disorder diagnosis)	-0.0289	0.9715	-3% odds of optimal sleep
N3Q20F - (Rape victims)	-0.0304	0.9700	-3% odds of optimal sleep
N3Q65A2 - (ADHD diagnosis)	-0.0527	0.9487	-5% odds of optimal sleep
N3Q65A35 - (Sleep apnea diagnosis)	-0.0649	0.9372	-6% odds of optimal sleep
N3Q20G - (Stalking victims)	-0.0746	0.9281	-7% odds of optimal sleep
RULS3 - (Loneliness scale score)	-0.0870	0.9166	-8% odds of optimal sleep
N3Q65A28 - (Insomnia diagnosis)	-0.1631	0.8495	-15% odds of optimal sleep
N3Q48 - (Perceived stress level)	-0.2956	0.7441	-26% odds of optimal sleep

A decision tree model was also tested for this classification idea, as sometimes decision trees may be able to capture nonlinear patterns otherwise not represented with logistic regression. This is due to decision trees having the capability of dividing the task into many sub-tasks that are addressed in a simple manner. The decision tree model resulted in similar but slightly less compelling metrics compared to the logistic regression model. The overall accuracy for the decision tree was 60.4% compared to 61.5% for the logistic regression model. Precision was 76.8% compared to 80.5% in the logistic regression model. Recall was nearly identical with the decision tree being 62.1% compared to 62.3%. The F1-score of the decision tree was 68.7% compared to 70.3% for the linear regression model. The specificity of the decision tree was 56.7% compared to 59.2%. Lasso, a statistical method that is used to prevent overfitting, was also tested. This, however, did not improve results.

A random forest model was also implemented. Random forests utilize multiple decision trees which may reduce variance and, in some cases, achieve better performance than just a single decision tree. However, the random forest actually produced metrics that were all weaker than the logistic regression model, and mostly equal to or weaker to the decision tree model. Notably, the accuracy was 60.1% and the ROC-AUC score was 0.625. An interesting caveat compared to the other models, however, is that higher resilience had roughly a 15% chance increase to increase the odds of optimal sleep compared to the 4% from the logistic regression model, whereas, conversely, the odds of optimal sleep for those with an anxiety diagnosis decreased to 5% from the 14% in the linear regression model. The effect from insomnia also decreased to around 6.5% from the 15% in the logistic regression. Stress, overall health, and resilience were the most impactful metrics in this model, each with over double the impact of any of the other predictors. This reinforces the notion of stress and overall health having larger impacts on sleep. Refer to Appendix E: Logistic Regression Modeling for the Python code used for this section of the pipeline. The next section of this report will discuss results and conclusions drawn from the review of the literature, data analysis, and modeling.

4 Results

Before discussing the implications of this study's findings, the results themselves are concisely summarized here. Data cleaning and analysis were conducted using the Python modules NumPy, Pandas, Scikit-Learn, and Statsmodels. The sample size of this data was 103,639 students. In the preliminary data analysis, correlations were made between sleep duration and DIENER well-being score, GPA, overall health, rape victimization, resilience, sexual assault victimization, stalking victimization, stress, time spent partying per week, and other variables. Additional correlations were made between the time needed to fall asleep and overall health rating, as well as the comorbidity of sleep apnea and insomnia among other mental and behavioral diagnoses. These correlations would be further investigated in the later modeling process.

After cleaning and preliminary analysis, a linear regression model was developed, but it resulted in insignificant findings. Next, a logistic regression model was developed, resulting in notable findings. The cleaning of the model was simply the removal of the data from students who had reported over 9 hours of sleep. The sample size after cleaning for this model was 102,173 students. Of these 102,173 students, 57,661 students (56.4%) were classified as achieving optimal sleep (7-9 hours per night), while the other 44,512 (43.6%) were classified as suboptimal sleepers (6 hours or less). The sleeping data used here was self-reported by the students and was based on their last two weeks of weeknight sleep. This logistic regression model predicting optimal versus suboptimal sleep was developed to examine the combined predictive power of psychological, diagnostic, behavioral, and trauma-related variables on sleep duration. Thirteen different predictors between these groups were used to attempt to predict optimal sleep or lack thereof. The model resulted in 61.5% accuracy and a ROC-AUC score of 0.637.

Stress (N3Q48) had the strongest negative effect on sleep (odds ratio = 0.74). The presence of loneliness had a considerable effect on decreasing the chances of optimal sleep, comparatively (odds ratio = 0.916). Diagnosis of insomnia (N3Q65A28) showed the strongest negative effect on sleep among all diagnoses (odds ratio = 0.85). Diagnosis of anxiety unexpectedly raised the chances of optimal sleep (odds ratio = 1.14). Better overall health ratings (N3Q1) increased the odds of optimal sleep (odds ratio = 1.22). Victims of sexual assault, rape, and stalking were all less likely to get optimal sleep (sexual assault victimization odds ratio = 0.982, rape victimization odds ratio = 0.970, stalking victimization odds ratio = 0.928). Refer to Table 7 above to view the summary of odds ratios and interpretations from the logistic regression model. The next section will discuss the interpretation of this study's results and compare findings to those in the review of the literature.

5 Discussion

The results of the logistic regression model, while not all-encompassing or extremely powerful, still indicates moderate predictive power, with the predictors being meaningful and informative in explaining the sleep duration of college students. The logistic regression model predicting optimal versus suboptimal sleep resulted in 61.5% accuracy and a ROC-AUC score of 0.637. The model correctly identified 81% of optimal sleepers. Linear regression, decision tree, and random forest machine learning models were also tested, but logistic regression produced the best results. The following paragraphs will discuss the interpretation of the results from the model as well as the preliminary data analysis. Sleep quality is multidimensional; it is affected but psychological, emotional, mental, and environmental conditions. Maintaining and regulating each of these factors, and being consistent in daily lifestyle choices, has shown a strong connection with optimal sleep quality [8].

Stress was identified in the preliminary data analysis as a moderately strong predictor for weeknight sleep duration. In the data analysis and visualization section, Figure 11 showed that the students who slept most optimally were those who had low stress rather than the group with no stress, which was somewhat unexpected. After utilizing the logistic regression model, results showed that perceived stress level of students was the strongest negative contributor and the single strongest contributor in the entire model, where each tier of stress (no/low/moderate/high stress) decreased the odds of optimal sleep by 26%. Interestingly, this causes a slight disparity between the preliminary findings and the model findings, where the model indicates that having no stress is ideal for optimal sleep rather than low stress. This minor discrepancy may be a result of the logistic regression model capturing the overall trend, where higher stress levels are identified as being more harmful. In reality, stress and performance are not entirely linear, and some minor stress may support motivation and routine in people. Regardless, stress is the primary risk factor in this model. Stress has been shown to be a common contributor to poor sleep habits [5] and has also been shown to increase as a result of suboptimal sleeping habits, shown especially in students with insomnia [10]. Stress has also been shown to be the primary risk factor for poor academic performance [2]. Therefore, the findings on stress in this study align with those in the review of the literature.

On the opposite end of the model from stress was the self-reported overall health ratings of the students. As the health ratings of the students improved (poor/fair/good/very good/excellent), the odds of optimal sleep improved by 22%. This reflects the general sentiment expressed in much of the literature studied as well as in the modeling results for this research project, which indicates that there are a multitude of factors that affect sleep quality, some of which are more easily controlled for than others, but those who generally have better health will generally have more optimal sleep than those with worse overall health. The preliminary data analysis showed a consistent increase in the number of students who received 7 or 8 hours of sleep on average as overall health rating increased, whereas every sleep duration less than 7 hours appeared less as the health ratings improved. For students with 9 hours of sleep on average, the proportions stayed roughly equal.

Resilience was defined in this survey as a student's ability to bounce back from illness, injury, or hardship. The model results indicate that the odds of optimal sleep increase by a modest 4% with each level of resilience (not/rarely/sometimes/often/very often resilient), indicating that students who possess stronger coping and recovery skills tend to sleep slightly more optimally. In the random forest model, however, this number higher was 15%. In the preliminary analysis, resilience did portray a consistent relationship with opt imal sleep duration as resilience responses improved. As the resilience responses improved, the number of students who reported

5 or fewer hours of sleep decreased consistently, the number of students who reported 6 hours of sleep stayed roughly the same, and the number of students who reported 7 or 8 hours of sleep increased quite consistently. This similarly reflects the findings of the students' overall health ratings, but with a much more nuanced effect. Figures 8 and 9 from the data analysis and visualization section visualize this relationship.

The presence of loneliness, as identified by the ULS3 loneliness assessment, predicted a decrease in optimal sleep by 8%. Loneliness has shown strong association with academic impairment and sleep difficulties, and it has been recommended that students avoid social isolation for that reason [4], likely among many other reasons.

Six diagnoses were tracked in the modeling and analysis of this research. These diagnoses were anxiety, depression, substance use, ADHD, sleep apnea, and insomnia. The presence of insomnia had the strongest diagnosis-related effect, predicting 15% lower odds of optimal sleep. This is logical, as insomnia is a sleep disorder. The effect of sleep apnea was only 6% lower odds of predicting optimal sleep. In the analysis, insomnia was found to have comorbidity with each of these disorders as well, with substance abuse and sleep apnea as the most common, but with anxiety and depression as the most proportionately significant. For example, 2.37% of students without a depression diagnosis reported having insomnia, whereas 20.34% of students with a depression diagnosis reported having insomnia as well, over 8 times as many. There were around 3 times more students who had insomnia and another diagnosis than there were students who had sleep apnea and another diagnosis, proportionately speaking. Students with insomnia and at least one other diagnosis included (anxiety, depression, substance use, ADHD, sleep apnea) also reported having academic issues around 2-3 times more than those with only insomnia. Obstructive sleep apnea and insomnia have put students at higher risk of academic shortcomings [21]. Chronic insomnia specifically has been shown to correlate with lower total sleep time, lower sleep efficiency, increased mental and physical fatigue, elevated anxiety, elevated depression, an increase in perceived stress and an overall lower quality of life. [10]. Insomnia has also been correlated with a greater risk of suicide [16], a greater risk of substance abuse, and a more significant number of reports of medical problems [10]. People who have dealt with trauma are at a greater risk of developing insomnia, seen especially in those who have experienced childhood trauma [9].

The presence of anxiety unexpectedly increased the odds of optimal sleep by 14%, while depression also raised the odds by 3%. The presence of substance use disorder slightly lowered the odds of optimal sleep at 3%, and the presence of ADHD and sleep apnea also lowered the odds by 5% and 6% respectively. The anxiety and depression findings diverge somewhat from the research reviewed for this study, although the findings on the other diagnoses were mostly parallel with previous studies.

As has already been well established, ADHD, sleep apnea, and substance use are all associated with poor sleep quality, with higher rates of sleep problems among people with one of these diagnoses than among those without [2] [1]. Anxiety and depression have also been understood to be not just contributors, but also consequences of poor sleep quality in previous studies, similar to many of these other contributors [7]. For example, higher severity anxiety has been linked to reduced sleep quality, while cognitive emotional regulation strategies have been suggested to improve sleep quality, mental and sleep disorder symptoms, and therefore overall quality of life [20].

Research has further demonstrated that extreme and undesired ADHD and depression symptoms were also correlated with worse sleep quality [19]. Anxiety and substance use are also closely related to sleep problems, especially among people with trauma, particularly those with childhood trauma. [9]. Depression and sleep have been observed to have a bidirectional relationship, as they both influence each other's behavior. Sleep quality seems to be more consistently a predictor of depressive factors than the opposite, however [17]. In particular, depression has been shown to be more common in people who regularly obtain less than 7 hours of sleep per night [23], and students with elevated symptoms of ADHD have been shown to experience significantly worse sleep [18]. Finally, alcohol and marijuana use have been associated with sleep difficulties [4], and more specifically, nighttime alcohol use has been shown to be common in those with sleeping disorders [21].

The final set of measured predictors were trauma-related predictors, those being victimization of sexual assault, rape, and stalking, within the last year. Each of the contributors decreased the odds for optimal sleep, with stalking victimization being the strongest at 7% lower odds, followed by rape victimization at 3% lower odds and sexual assault victimization at 2% lower odds. As discussed with the other predictors, childhood trauma has been shown to interfere with quality sleep, and thus it is reasonable to infer that more newly developed trauma could also do the same [9]. Exposure to traumatic experiences disturbs sleep.

In general, many findings were roughly as expected. The presence of anxiety increased the chances of optimal sleep, which was unexpected, as did the chances of optimal sleep increasing for those with depression, although to a more minor degree. This could reflect coping mechanisms, treatment effects, or measurement overlap, such as anxious students who are more aware of sleep hygiene. Although the literature did suggest an undesirable relationship between anxiety and sleep and depression and sleep as contributors to and consequences of poor sleep quality, most of the focus was on consequential outcomes, while this modeling focused on contributions to sleep quality, which may also help explain this. Also, although the linear regression model had relatively poor performance, it is worth noting that depression was shown to be strongly associated with reduced sleep in that model, unlike the logistic regression model, whereas anxiety was shown to predict slightly higher sleep either way. The final main section of the written report will summarize the overall findings and implications of this research and highlight their relevance to existing research and as potential topics for future research.

6 Conclusion

The purpose of this study at the beginning was to investigate the factors that correlate with sleep quality among college students, but as the research process continued, the focus narrowed to investigating what factors contribute to sleep optimality and sub-optimality among college students, specifically using the ACHA-NCHA III spring 2024 dataset. Preliminary data analysis was conducted using stacked bar charts, heatmaps, barplots, and other visualization methods. The preliminary data analysis helped inform the modeling process, where a linear regression model was tested to predict sleep outcomes, but performed worse than desired. After this, a logistic regression model was developed to classify optimal and suboptimal sleepers, using many of the predictors used in the linear regression model. The suboptimal group and the optimal group were divided based on findings of other studies addressed in the review of the literature, as well as the findings in the preliminary data analysis.

The key findings of this study include that stress had the strongest negative relationship with optimal sleep and that insomnia has strong patterns of comorbidity with substance abuse, sleep apnea, ADHD, depression, and anxiety, in addition to having a strong effect on the model. College students with insomnia often also struggle with academics. The better overall health scores had the strongest positive relationship with optimal sleep. The model achieved an accuracy of 61.5% and had moderate predictive power. These findings align with many of the results discussed in the review of the literature and generally support the role stress and mental health play in poor sleep. A key difference from the findings of the other studies that was unexpected was that anxiety had a positive relationship with optimal sleep, which could indicate behavioral or coping differences. Although the findings in this study are not necessarily groundbreaking, the vast majority of the findings add support to what has previously been studied on the topic. The findings on stress, trauma, and insomnia from this study add some extra nuance to previous research and outline the potential for future studies that are more focused on specific contributors, such as trauma and insomnia.

This topic and these findings are important to college students, institutions, researchers, and every other human being alike. Sleep is extremely important for every human being and investigating what contributes to optimal and suboptimal sleep to inform other people how to get the most out of their sleep is important, as sleep is a catalyst that plays off of the inputs it receives and the outputs it gives in a cyclic fashion. For college students specifically, these results matter, as they add more credibility to the importance of stress management, sleep hygiene education, and awareness of comorbid conditions and their effects on sleep. These results and this subject matter are also important for institutions, as they help institutions provide mental health support more effectively and develop prevention programs to help people overcome their health problems. For researchers, this study is valuable as it both adds a little to the conversation about sleep and also supports existing data on the topic. For future researchers using the ACHA-NCHA spring 2024 dataset or other college health datasets, this is especially valuable.

Some limitations that exist with this study are the cross-sectional nature of the ACHA-NCHA data, as the focus is not on the sleep of college students, but is instead focused on the general health of college students and the multitude of factors that contribute to that. Due to this and also the majority of the data being self-reported, there is a lack of causal inference here. Another weakness is that the model did not include students who reported an average of more than 9 hours of sleep, which was a deliberate choice for the sake of creating a simple model and due to somewhat conflicting data and opaque reasoning for high amounts of sleep. Unfortunately, however, no data for students who would be in the range of 6-7 hours was included because of the way the survey data was organized. If the reported amounts of sleep were more specific and this common subtype of sleep duration was accessible, the model is expected to have performed more accurately. The model may also have benefited if it could read not only the duration of weeknight sleep, but also the duration of weekend sleep. One other small weakness here is that the dataset did not include variables on caffeine, screen time, and some other effects that contribute to sleep quality that would have been included in this study if included in the data. For future studies, other contributors should be investigated. These include, but are not limited to, caffeine usage, screen time, more specific questions on alcohol use, other diagnoses, and also academic year, which was briefly discussed in demographics but not utilized in the analysis. More advanced and varied models should be used to test causality as well, especially with longitudinal data and data with a narrower focus. Differences between subgroups should also be investigated further, such as the differences between gender, athletes and non-athletes, and students who live in Greek Life housing or do not, for example.

Overall, this study reinforces the strong influence that psychological and diagnostic factors have on sleep quality in college students. This study also reveals the importance of approaches and interventions to improve overall well-being and sleep quality.

7 Acknowledgements

First, I would like to thank my parents for supporting me personally, financially and academically, always pushing me to strive for excellence.

This research project utilized data from the American College Health Association National College Health Assessment III (ACHA-NCHA III) spring 2024 dataset. I would like to thank the ACHA for allowing me to use their data to investigate this topic.

I would like to thank Adrian College for providing quality learning opportunities, such as my capstone class, where I was able to work with two friends and receive direct support from my supervisor. Speaking of my capstone course supervisor, I would like to extend my gratitude to Dr. Yasser Alginahi, who guided me along the way in developing this project and offered his support and time, pushing me out of my comfort zone when I needed it.

I have also enjoyed the company of my colleagues, Seth Logan and Anthony Mollica. We all worked together on our projects, though each of us researched separate things, and they, alongside Dr. Alginahi, offered me advice on many aspects of this project.

References

[1] S. P. Becker, M. A. Jarrett, A. M. Luebbe, A. A. Garner, G. L. Burns, and M. J. Kofler, "Sleep in a large, multi-university sample of college students: sleep problem prevalence, sex differences, and mental health correlates," *Sleep health*, vol. 4, no. 2, pp. 174–181, 2018.

[2] P. Upright, T. Esslinger, and W. Hays, "Health issues affecting college student's academic performance," *Kentucky Association of Health, Physical Education, Recreation and Dance*, vol. 51, no. 2, pp. 30–36, 2014.

[3] A. M. Lederer, S. B. Oswalt, M. T. Hoban, and M. N. Rosenthal, "Health-related behaviors and academic achievement among college students," *American journal of health promotion*, vol. 38, no. 8, pp. 1129–1139, 2024.

[4] S. Chung, "Academic impairment from sleep difficulties: The role of substance use, psychological distress, and loneliness in us college students," *medRxiv*, pp. 2025–06, 2025.

[5] T. Deliens, B. Deforche, I. De Bourdeaudhuij, and P. Clarys, "Determinants of physical activity and sedentary behaviour in university students: a qualitative study using focus group discussions," *BMC public health*, vol. 15, no. 1, p. 201, 2015.

[6] K. Peltzer and S. Pengpid, "Nocturnal sleep problems among university students from 26 countries," *Sleep and Breathing*, vol. 19, no. 2, pp. 499–508, 2015.

[7] S. D. Hershner and R. D. Chervin, "Causes and consequences of sleepiness among college students," *Nature and science of sleep*, pp. 73–84, 2014.

[8] N. Zisapel, "New perspectives on the role of melatonin in human sleep, circadian rhythms and their regulation," *British journal of pharmacology*, vol. 175, no. 16, pp. 3190–3199, 2018.

[9] L. D. Albers, T. J. Grigsby, S. M. Benjamin, C. J. Rogers, J. B. Unger, and M. Forster, "Adverse childhood experiences and sleep difficulties among young adult college students," *Journal of sleep research*, vol. 31, no. 5, p. e13595, 2022.

[10] D. J. Taylor, A. D. Bramoweth, E. A. Grieser, J. I. Tatum, and B. M. Roane, "Epidemiology of insomnia in college students: relationship with mental health, quality of life, and substance use difficulties," *Behavior therapy*, vol. 44, no. 3, pp. 339–348, 2013.

[11] X. Liu, L. Lang, R. Wang, W. Chen, X. Ren, Y. Lin, G. Chen, C. Pan, W. Zhao, T. Li *et al.*, "Poor sleep quality and its related risk factors among university students," *Annals of palliative medicine*, vol. 10, no. 4, pp. 4 479 485–4 474 485, 2021.

[12] D.-V. Phan, C.-L. Chan, R.-H. Pan, N.-P. Yang, H.-C. Hsu, H.-W. Ting, K. R. Lai, and K.-B. Lin, "Investigating the effect of daily sleep on memory capacity in college students," *Technology and Health Care*, vol. 27, no. 2, pp. 183–194, 2019.

[13] "Working memory capacity — an overview," ScienceDirect Topics, accessed 2025-09-18, https://www.sciencedirect.com/topics/computer-science/working-memory-capacity.

[14] M. E. Patrick, J. Griffin, E. D. Huntley, and J. L. Maggs, "Energy drinks and binge drinking predict college students' sleep quantity, quality, and tiredness," *Behavioral sleep medicine*, vol. 16, no. 1, pp. 92–105, 2018.

[15] M. S. Mahfouz, S. A. Ali, A. Y. Bahari, R. E. Ajeebi, H. J. Sabei, S. Y. Somaily, Y. A. Madkhali, R. H. Hrooby, and R. N. Shook, "Association between sleep quality and physical activity in saudi arabian university students," *Nature and Science of Sleep*, pp. 775–782, 2020.

[16] K. Russell, S. Allan, L. Beattie, J. Bohan, K. MacMahon, and S. Rasmussen, "Sleep problem, suicide and self-harm in university students: A systematic review," *Sleep medicine reviews*, vol. 44, pp. 58–69, 2019.

[17] J. Dinis and M. Bragança, "Quality of sleep and depression in college students: A systematic review," *Sleep Science*, vol. 11, no. 4, pp. 290–301, 2018. [Online]. Available: https://pmc.ncbi.nlm.nih.gov/articles/PMC6361309/

[18] A. K. Demirkan, B. Yildirim, F. Ozdemir, E. Sari, and C. Kose, "Investigating adhd symptoms and sleep disturbances in young adults: A cross-sectional study," *BMC Psychology*, vol. 13, p. 112, 2025. [Online]. Available: https://pmc.ncbi.nlm.nih.gov/articles/PMC12266080/

[19] J. Kim, E. H. Hwang, S. Shin, and K. H. Kim, "University students' sleep and mental health correlates in south korea," *Healthcare*, vol. 10, no. 8, p. 1515, 2022. [Online]. Available: https://www.researchgate.net/publication/363057912_University_Students%27_Sleep_and_Mental_Health_Correlates_in_South_Korea

[20] Y. Wang, Z. Guang, J. Zhang, L. Han, R. Zhang, Y. Chen, Q. Chen, Z. Liu, Y. Gao, R. Wu, and S. Wang, "Effect of sleep quality on anxiety and depression symptoms among college students in china's xizang region: The mediating effect of cognitive emotion regulation," *Behavioral Sciences*, vol. 13, no. 10, p. 861, 2023. [Online]. Available: https://www.mdpi.com/2076-328X/13/10/861

[21] A. Jain and S. Verma, "Prevalence of sleep disorders among college students: a clinical study," *Journal of Advanced Medical and Dental Sciences Research*, vol. 4, no. 6, p. 103, 2016.

[22] A. H. Dockery, "Examining sleep and sleep hygiene in a sample of college students and differences between on and off-campus housing," Ph.D. dissertation, University Honors College, Middle Tennessee State University, 2022.

[23] C. C. Panel, N. F. Watson, M. S. Badr, G. Belenky, D. L. Bliwise, O. M. Buxton, D. Buysse, D. F. Dinges, J. Gangwisch, M. A. Grandner *et al.*, "Recommended amount of sleep for a healthy adult: a joint consensus statement of the american academy of sleep medicine and sleep research society," *Journal of Clinical Sleep Medicine*, vol. 11, no. 6, pp. 591–592, 2015.

[24] C. Clinic. (n.d.) Sleep apnea. Accessed: 2025-10-19. [Online]. Available: https://my.clevelandclinic.org/health/diseases/8718-sleep-apnea

[25] WebMD. (n.d.) Insomnia: Symptoms and causes. Accessed: 2025-10-19. [Online]. Available: https://www.webmd.com/sleep-disorders/insomnia-symptoms-and-causes

[26] Wikipedia contributors, “Linear regression,” https://en.wikipedia.org/wiki/Linear_regression, accessed: 2025-11-16.

[27] M. Vandekerckhove and R. Cluydts, “The emotional brain and sleep: An intimate relationship,” *Sleep Medicine Reviews*, vol. 14, no. 4, pp. 219–226, 2010.

[28] L. Author and A. Author, “Title of the article,” *BioMedica*, vol. 5, no. –, p. –, 2022.

[29] Wikipedia contributors, “Logistic regression,” https://en.wikipedia.org/wiki/Logistic_regression, accessed: 2025-11-16.

[30] M. Perlis, P. J. Shaw, G. Cano, and C. A. Espie, *Models of Insomnia.* Elsevier, 2010, pp. 850–865.

[31] D. Riemann, K. Spiegelhalder, B. Feige, C. Voderholzer, K. Hornyak, U. Nissen, F. Hennig, and C. Baglioni, “The hyperarousal model of insomnia: A review of the concept and its evidence,” *Sleep Medicine Reviews*, vol. 14, no. 1, pp. 19–31, 2010.

Appendices

7.1 Appendix A: Cleaning

```
’’’
Jonah Watson
Fall 2025
Sleep Health in College Students: A Multivariable Predictive Modeling Analysis
This file was used to clean the ACHA-NCHA Spring 2024 survey data used in this
project
’’’

import pandas as pd             # imported for data manipulation and analysis



# open the original CSV file
df = pd.read_csv("SLEEP NCHA-III S24 - New_Numeric.csv")

# keep specific columns (removing unnecessary variables)
keep_variables = [
    ’N3Q1’, ’N3Q3E’, ’N3Q3I’, ’N3Q13’, ’N3Q14’, ’N3Q15’, ’N3Q16A’, ’N3Q16B’,
    ’N3Q16C’, ’N3Q16D’, ’N3Q16E’, ’N3Q20D’, ’N3Q20F’, ’N3Q20G’, ’N3Q41C’,
    ’N3Q42B’, ’N3Q48’, ’N3Q65A2’, ’N3Q65A3’, ’N3Q65A7’, ’N3Q65A15’, ’N3Q65A28’,
    ’N3Q65A35’, ’N3Q65Y’,’N3Q66P’, ’N3Q67A’, ’N3Q69’, ’N3Q72’, ’N3Q75A1’,
    ’N3Q75A2’, ’N3Q75A3’, ’N3Q75A4’, ’N3Q75A5’, ’N3Q75A6’, ’N3Q75A7’,
    ’N3Q75A8’, ’N3Q77B’, ’N3Q80’, ’RBMI’, ’RKESSLER6’, ’RULS3’, ’DIENER’]

# save the variables I want to keep for analysis
df = df[keep_variables]



"""This next section shows the process of recoding responses for certain
variables to reduce the number of categories and allow for a more easily
digestable analysis. All data had already been converted to numeric at this
point"""

def recode_NQ3E_I(x):
    """The variables N3Q3E and N3Q3I represent a student’s time spent doing
    physical activity and time spent partying per week, respectively. This
    function recodes the categories into larger bins"""
    if x in [1, 2]:         # 0-5 hours
        return 1
    elif x in [3, 4]:       # 6-15 hours
        return 2
    elif x in [5, 6]:       # 16-25 hours
        return 3
    elif x in [7, 8]:       # 26+ hours
```

```
        return 4
    return pd.NA            # if the response does not belong

def recode_NQ13(x):
    """The variable N3Q13 represents how long it takes a student to fall
    asleep. This function recodes the categories into larger bins"""
    if x in [1, 2]:         # <15 minutes
        return 1
    elif x == 3:
        return 2            # 16-30 minutes
    elif x in [4, 5]:       # >30 minutes
        return 3
    return pd.NA

def recode_NQ16A_E(x):
    """The variables N3Q16A, N3Q16B, N3Q16C, N3Q16D, and N3Q16E represent
    questions that relate to a student's sleeping habits, based on the last
    7 days. This function recodes the categories into larger bins. For more
    specific information on the questions that these variables represent,
    please see the Data Collection and Profiling subsection under Methodology
    in the research report."""
    if x in [1, 2]:         # 0-1 days
        return 1
    elif x in [3, 4]:       # 2-3 days
        return 2
    elif x in [5, 6, 7, 8]: # 4+ days
        return 3
    return pd.NA

def recode_NQ41C(x):
    """The variable N3Q41C represents a student's response to if they feel
    that they are engaged and interested in their daily activites. This
    function recodes the categories into larger bins"""
    if x in [1, 2, 3]:     # disagree
        return 1
    elif x == 4:           # neither
        return 2
    elif x in [5, 6, 7]:   # agree
        return 3
    return pd.NA

def recode_RBMI(x):
    """The variable RBMI represents a student's body mass index. This
    function recodes the categories into larger bins"""
    if x == 1:             # underweight
        return 1
    elif x == 2:           # healthy weight
        return 2
    elif x == 3:           # overweight
        return 3
    elif x in [4, 5, 6]:   # obese
```

```
        return 4
    return pd.NA



# apply recodes to dataframe
df["N3Q3E_recode"] = df["N3Q3E"].apply(recode_NQ3E_I)
df["N3Q3I_recode"] = df["N3Q3I"].apply(recode_NQ3E_I)
df["N3Q13_recode"] = df["N3Q13"].apply(recode_NQ13)

# for loop for recoding these 5 variables since they are in an iterable format
for var in ["N3Q16A", "N3Q16B", "N3Q16C", "N3Q16D", "N3Q16E"]:
    df[f"{var}_recode"] = df[var].apply(recode_NQ16A_E)

df["N3Q41C_recode"] = df["N3Q41C"].apply(recode_NQ41C)
df["RBMI_recode"] = df["RBMI"].apply(recode_RBMI)



# save the cleaned data to a new CSV file
df.to_csv("CLEANED SLEEP NCHA-III S24 - New_Numeric.csv", index=False)

print('Cleaned data saved')

'''
AI assistance was used to suggest solutions and resolve errors
'''
```

7.2 Appendix B: Profiling

```
'''
Jonah Watson
Fall 2025
Sleep Health in College Students: A Multivariable Predictive Modeling Analysis
This file was used to profile the ACHA-NCHA Spring 2024 survey data used in
this project.
'''

import pandas as pd  # imported for data manipulation and analysis
import json   # imported to access the JSON file storing column label mappings



# load the cleaned dataset
df = pd.read_csv("CLEANED SLEEP NCHA-III S24 - New_Numeric.csv")

# load the JSON label mappings
with open("NCHA-IIIb_labels_S24_copy.json", 'r') as jsonfile:
    labels = json.load(jsonfile)



'''Prepare a text file to store the profiling data and response rates of each
choice for each survey question'''
response_rates = "sleep_study_response_rates.txt"

with open(response_rates, 'w') as f:
    f.write("Number of rows and columns: \n\n")
    f.write(str(df.shape) + "\n\n")

    f.write("General information about the data: \n\n")
    df.info(buf=f)
    # buf=f is used redirect df.info() right into the text file
    f.write("\n\n")

    f.write("Quick summary statistics: \n\n")
    f.write(str(df.describe().T) + "\n\n")
    # .T to flip rows and columns (I find it easier to read)

    f.write("Missing values per column: \n\n")
    f.write(str(df.isnull().sum()) + "\n\n")

    f.write("Count of duplicate rows: \n\n")
    f.write(str(df.duplicated().sum()) + "\n\n")
    df = df.drop_duplicates()
    f.write("Duplicate rows have now been dropped\n\n")
    f.write("Count of duplicate rows is now: \n\n")
    f.write(str(df.duplicated().sum()) + "\n\n")
    f.write("Value counts per column: \n\n")



    for column in df.columns:
        """This for loop counts the number of responses per column in the
        dataset by looping through every column in the dataset, counting
        the frequency of each unique value (including missing values),
        and matching numeric codes to descriptive labels in the JSON file,
        where the results are then written to a text file."""
        f.write(f"\n\n--- {column} ---\n\n")
        # write the columns name as a section header in the text file

        counts = df[column].value_counts(dropna=False).sort_index()
        # count how many times unique values appear in each column
        # dropna=False included to also count missing values

        variable_name = column.split(":")[0].strip()
        # this is the appropriate way to match variable names based on the way...
        # ...I renamed columns in my dataframe
        # example:  "N3Q1": "N3Q1: Overall health rating" becomes "N3Q1"

        if variable_name in labels:
        # my JSON file (labels) contains all of the data prior to cleaning,...
        # ...so some labels won’t be used (that’s why I use "if")

            for code, count in counts.items():
            # for the JSON labels associated with the variables that I am using...
            # ...count each response rate

                if pd.isna(code):
                    label = "Missing"
                else:
                    label = labels[variable_name].get(str(int(code)), code)
                # if the data is blank, assign it "Missing" in the text file
                # otherwise, look at the JSON for what the numeric correlation is
                # example. in some questions "1" = "Poor", "2" = "Fair", etc.

                f.write(f"{label} ({code}): {count}\n\n")
                # write the results into the text file
        else:
            f.write(str(counts) + "\n\n")
            # if for some reason, the columns aren’t able to be matched with the...
            # ...JSON file, still print them

print("Data successfully written to text file")

’’’
AI assistance was used to suggest solutions and resolve errors
’’’
```

7.3 Appendix C: Analyzing

```
'''
Jonah Watson
Fall 2025
Sleep Health in College Students: A Multivariable Predictive Modeling Analysis
This file was used to profile the ACHA-NCHA Spring 2024 survey data used in
this project.
'''

import pandas as pd           # imported for data manipulation and analysis
import seaborn as sns         # imported for data visualization
import matplotlib.pyplot as plt  # imported for data visualization
import numpy as np              # imported to perform operations on arrays



# load the cleaned dataset
df = pd.read_csv("CLEANED SLEEP NCHA-III S24 - New_Numeric.csv")



# sleep variables (predictors)
sleep_vars = ["N3Q13_recode", "N3Q14", "N3Q15"]
# the following list was also used with "sleep_vars" before the impact of the...
# ...variables within were found to be negligible:
# ["N3Q16A_recode", "N3Q16B_recode", "N3Q16C_recode", "N3Q16D_recode", "N3Q16E_recode"]

# demographic variables
demographic_vars = ["N3Q67A", "N3Q72", "N3Q75A1", "N3Q75A2", "N3Q75A3",
"N3Q75A4", "N3Q75A5", "N3Q75A6", "N3Q75A7", "N3Q75A8", "N3Q77B", "N3Q80",
"RBMI_recode"]

# diagnosis variables
diagnosis_vars = ["N3Q65A2", "N3Q65A3", "N3Q65A7", "N3Q65A15", "N3Q65A28",
"N3Q65A35", "N3Q65Y"]

# sleep disorder variables
sleep_disorder_vars = ["N3Q65A28", "N3Q65A35"]

# outcome variables
outcome_vars = ["N3Q1", "N3Q3E_recode", "N3Q3I_recode", "N3Q20D", "N3Q20F",
"N3Q20G", "N3Q41C_recode", "N3Q42B", "N3Q48", "N3Q66P", "RKESSLER6",  "RULS3",
"DIENER"]



'''Prepare a text file to store the correlations between sleep variables
and demographic variables to be utilized'''
sleep_demographic_file = "sleep_demographic_correlations.txt"
```

```
"""The following block of code underneath the "with" statement will be reused
and slightly modified several more times for each of the different variable
classes that were created above."""
with open(sleep_demographic_file, 'w') as f:
    for demographic in demographic_vars:
        for sleep_var in sleep_vars:

            # write the percentages within each demographic
            percentages = pd.crosstab(df[sleep_var], df[demographic],
                                      normalize='columns') * 100
            f.write(f"\n{sleep_var} vs {demographic}\n")
            f.write(percentages.round(2).to_string())
            # converting to string because the data is being written to a file
            f.write("\n\n")

            # write the mean value of demographic variables for each category...
            # ...of sleep variable
            means = df.groupby(demographic)[sleep_var].mean()
            f.write(f"\nMean {sleep_var} per {demographic}:")
            f.write(percentages.round(2).to_string())
            f.write("\n\n" + "="*40 + "\n")         # divider for formatting
    print(f"Correlation analysis saved to {sleep_demographic_file}\n\n")



'''Calculate and visualize the correlations between different GPAs and time
taken to fall asleep'''
means = df.groupby('N3Q80')['N3Q13_recode'].mean()
means.plot(kind='bar', title='Mean of Time to Fall Asleep by GPA',
           color='mediumslateblue')

plt.ylim(1, 3)
y_bins = ("0-15 minutes", "16-30 minutes", "31+ minutes")
plt.yticks([1, 2, 3], y_bins)
plt.ylabel("Mean Time to Fall Asleep")

x_bins = ['A+', 'A', 'A-', 'B+', 'B', 'B-', 'C+', 'C', 'C-', 'D+', 'D',
          'D-', 'F']
plt.xticks(ticks=range(len(x_bins)), labels=x_bins, rotation=0)
plt.xlabel("GPA")

plt.tight_layout()
plt.savefig("mean_time_to_fall_asleep_by_GPA.png")



'''Calculate and visualize the correlations between different GPAs
and average weeknight asleep'''
means = df.groupby('N3Q80')['N3Q14'].mean()
print(f"Means: {means}")
means.plot(kind='bar', title='Average Weeknight Sleep by GPA',
           color=’mediumslateblue’)

y_bins = ("<4 hours", "4 hours", "5 hours", "6 hours", "7 hours", "8 hours",
          "9 hours", ">9 hours",)
plt.yticks([1, 2, 3, 4, 5, 6, 7, 8], y_bins)
plt.ylabel("Average Weeknight Sleep")

x_bins = [’A+’, ’A’, ’A-’, ’B+’, ’B’, ’B-’, ’C+’, ’C’, ’C-’, ’D+’, ’D’,
          ’D-’, ’F’]
plt.xticks(ticks=range(len(x_bins)), labels=x_bins, rotation=0)
plt.xlabel("GPA")

plt.tight_layout()
plt.savefig("mean_weeknight_sleep_by_GPA.png")




’’’Calculate the correlations between sleep variables and diagnosis variables’’’
sleep_diagnosis_file = "sleep_diagnosis_correlations.txt"

with open(sleep_diagnosis_file, ’w’) as f:
    for diagnosis in diagnosis_vars:
        for sleep_var in sleep_vars:
            percentages = pd.crosstab(df[sleep_var], df[diagnosis],
                                      normalize=’columns’) * 100
            f.write(f"\n{sleep_var} vs {diagnosis}\n")
            f.write(percentages.round(2).to_string())
            f.write("\n\n")

            means = df.groupby(diagnosis)[sleep_var].mean()
            f.write(f"\nMean {sleep_var} per {diagnosis}:")
            f.write(means.round(2).to_string())
            f.write("\n\n" + "="*40 + "\n")
    print(f"Correlation analysis saved to {sleep_diagnosis_file}\n\n")



’’’Calculate the correlations between sleeping disorders (sleep apnea and
insomnia) and outcome variables’’’
sleep_disorder_file = "sleep_disorder_correlations.txt"

with open(sleep_disorder_file, ’w’) as f:
    for outcome in outcome_vars:
        for sleep_var in sleep_disorder_vars:
            percentages = pd.crosstab(df[sleep_var], df[outcome],
                                      normalize=’columns’) * 100
            f.write(f"\n{sleep_var} vs {outcome}\n")
            f.write(percentages.round(2).to_string())
            f.write("\n\n")
            means = df.groupby(outcome)[sleep_var].mean()
            f.write(f"\nMean {sleep_var} per {outcome}:")
            f.write(means.round(2).to_string())
            f.write("\n\n" + "="*40 + "\n")
    print(f"Correlation analysis saved to {sleep_disorder_file}")



'''Calculate the correlations between sleeping disorders (sleep apnea and
insomnia) and the included mental disorders’’’
dbd_file = "disorder_by_disorder_correlations.txt"

with open(dbd_file, ’w’) as f:
    for diagnosis in diagnosis_vars:
        for sleep_var in sleep_disorder_vars:
            percentages = pd.crosstab(df[sleep_var], df[diagnosis],
                                      normalize=’columns’) * 100
            f.write(f"\n{sleep_var} vs {diagnosis}\n")
            f.write(percentages.round(2).to_string())
            f.write("\n\n")

            means = df.groupby(diagnosis)[sleep_var].mean()
            f.write(f"\nMean {sleep_var} per {diagnosis}:")
            f.write(means.round(2).to_string())
            f.write("\n\n" + "="*40 + "\n")
    print(f"Correlation analysis saved to {dbd_file}")



'''Create an empty dictionary to count students that reported having certain
diagnoses (ADD/ADHD, anxiety, depression, sleep disorders, etc.)’’’
results = {}

for diagnosis in diagnosis_vars:
    if diagnosis == "N3Q65Y":
    # this variable has its own analysis & visualization
    # it wasn’t included in my "classes" above
        continue
    counts = df[diagnosis].dropna().value_counts(normalize=True) * 100
    counts = counts.reindex([1,2])  # responses were reindexed for consistency
    results[diagnosis] = counts

    '''Create a new dataframe from the results dictionary and fill
    missing values with 0’’’
    diagnosis_counts = pd.DataFrame(results).T.fillna(0)
    diagnosis_counts.columns = ["% who said No (1)", "% who said Yes (2)"]

    """Prepare a text file to store the information from survey question
    N3Q65Y (how many students out of the total survey size have
    been diagnosed with the disorders from the diagnosis_vars list created
    above)"""
    diagnosis_counts_file = "diagnosis_impacts.txt"

    with open(diagnosis_counts_file, 'w') as f:
        f.write("Percent of No (1) or Yes (2) for each category:\n\n")
        f.write(diagnosis_counts.round(2).to_string())
        f.write("\n\n" + "="*40 + "\n")

print(f"Diagnosis counts saved to {diagnosis_counts_file}\n\n")

# visualize the counts of each diagnosis
plot_df = diagnosis_counts.reset_index().rename(columns={'index': 'Diagnosis'})

x_labels = {
    "N3Q65A2": "ADD/ADHD",
    "N3Q65A3": "Alcohol/Drug Use Disorder",
    "N3Q65A7": "Anxiety",
    "N3Q65A15": "Depression",
    "N3Q65A28": "Insomnia",
    "N3Q65A35": "Sleep Apnea"
}
plot_df["Diagnosis"] = plot_df["Diagnosis"].replace(x_labels)

plt.figure(figsize=(6,6))
sns.barplot(data=plot_df, x='Diagnosis', y='% who said Yes (2)',
            color='#bd735b')

plt.title('Percentage of Students with Each Diagnosis (Yes Responses)')
plt.xticks(rotation=45, ha='right')
plt.ylabel('% Yes')
plt.ylim(0, 100)
plt.xlabel('Diagnosis Variable')
plt.tight_layout()
plt.savefig("students_by_diagnosis.png")



'''Calculate the count of responses for the question "N3Q66P: Last 12 months,
have sleep difficulties affected academic performance?'''
print("Response counts for N3Q66P: \n")
print("did not experience 1, experienced and no affect 2, experienced and " \
"bad class performance 3, experience - delayed degree 4")
df = df[df["N3Q66P"].notna()]
print(df["N3Q66P"].value_counts(normalize=True) * 100)



# calculate the correlations between sleep variables and outcome variables
sleep_outcome_file = "sleep_outcome_correlations.txt"

with open(sleep_outcome_file, 'w') as f:
    for outcome in outcome_vars:
        for sleep_var in sleep_vars:
            # percentages for each outcome
            percentages = pd.crosstab(df[sleep_var], df[outcome],
                                      normalize=’columns’) * 100
            f.write(f"\n{sleep_var} vs {outcome}\n")
            f.write(percentages.round(2).to_string())
            f.write("\n\n")

            # mean value of outcome variables for each sleep variable category
            means = df.groupby(outcome)[sleep_var].mean()
            f.write(f"\nMean {sleep_var} per {outcome}:")
            f.write(means.round(2).to_string())
            f.write("\n\n" + "="*40 + "\n")
    print(f"Correlation analysis saved to {sleep_outcome_file}")


# ===============
# Visualizations
# ===============


’’’Use a stacked bar chart to visualize the relationship between overall health
ratings and time taken to fall asleep. Note that the values for each
visualization come from the correlation text files created above’’’
health_labels = [’Poor’, ’Fair’, ’Good’, ’Very Good’, ’Excellent’]
fall_asleep_time = {
   "<15min": np.array([21.81,  28.36,  37.21,  46.68,  53.72]),
    "16-30min": np.array([20.04,  24.97,  28.97,  29.22,  26.80]),
    "31+min": np.array([58.15,  46.67,  33.82,  24.10,  19.47]),
}

width = 0.6

fig, ax = plt.subplots(figsize=(10,6))
bottom = np.zeros(len(health_labels))

"""The following four lines of code create a stacked bar chart where each
category on the y-axis gets its own color and is stacked on top previous
categories. This format is used for each of the following stacked bar chart
visualizations as well"""
colors = plt.cm.summer(np.linspace(0, 1, len(fall_asleep_time)))
for i, (timerange, percent) in enumerate(fall_asleep_time.items()):
    ax.bar(health_labels, percent, width, label=timerange, bottom=bottom,
           color=colors[i])
    bottom += percent

ax.set_title(’Time Needed to Fall Asleep by Overall Health Rating’,
             fontsize=14, weight=’bold’)
ax.set_ylabel(’Percentage of Students (%)’, fontsize=12)
ax.set_xlabel(’Overall Health Rating’, fontsize=12)
ax.set_ylim(0, 100)
ax.legend(title=’Time to Fall Asleep’, loc=’upper left’,
          bbox_to_anchor=(1.05, 1))
plt.tight_layout()
plt.savefig(’time_to_fall_asleep_by_overall_health.png’)



'''Use a heatmap to visualize the relationship between overall health ratings
and time taken to fall asleep'''
df_sleep_health = pd.DataFrame(fall_asleep_time, index=health_labels)
df_sleep_health = df_sleep_health.T
# .T transposes the data to make durations rows, and health ratings columns

plt.figure(figsize=(9,6))
sns.heatmap(
    df_sleep_health,
    cmap="Greens",
    annot=True,            # Show values inside cells
    fmt=".1f",             # One decimal place
    cbar_kws={’label’: ’Percentage of Students (%)’}
)

plt.title(’Time Needed to Fall Asleep by Overall Health Rating’, fontsize=14,
           weight=’bold’)
plt.xlabel(’Overall Health Rating’, fontsize=12)
plt.ylabel(’Percentage of Students (%)’, fontsize=12)
plt.tight_layout()
plt.savefig(’heatmap_time_to_fall_asleep_by_overall_health.png’)


'''Use a stacked bar chart to visualize the relationship between overall
health ratings and weeknight sleep duration'''
health_labels = [’Poor’, ’Fair’, ’Good’, ’Very Good’, ’Excellent’]
weeknight_sleep_durations = {
    "<=4h": np.array([6.54, 2.72, 1.10, 0.59, 1.05]),
    "4h": np.array([8.79, 6.04, 3.45, 1.93, 2.19]),
    "5h": np.array([20.84, 17.68, 13.23, 8.68, 7.74]),
    "6h": np.array([27.32, 30.71, 28.77, 24.73, 22.18]),
    "7h": np.array([17.64, 25.08, 32.27, 36.66, 34.80]),
    "8h": np.array([11.17, 11.74, 16.15, 21.97, 26.04]),
    "9h": np.array([3.81, 3.83, 3.67, 4.37, 4.44]),
    ">9h": np.array([3.88, 2.21, 1.37, 1.06, 1.56]),
}

width = 0.6

fig, ax = plt.subplots(figsize=(10,6))
bottom = np.zeros(len(health_labels))

colors = plt.cm.nipy_spectral(np.linspace(0, 1, len(weeknight_sleep_durations)))
for i, (timerange, percent) in enumerate(weeknight_sleep_durations.items()):
    ax.bar(health_labels, percent, width, label=timerange, bottom=bottom,
           color=colors[i])
    bottom += percent

ax.set_title(’Average Weeknight Sleep Duration by Overall Health Rating’,
             fontsize=14, weight=’bold’)
ax.set_ylabel(’Percentage of Students (%)’, fontsize=12)
ax.set_xlabel(’Overall Health Rating’, fontsize=12)
ax.set_ylim(0, 100)
ax.legend(title=’Average Weeknight Sleep Duration’, loc=’upper left’,
          bbox_to_anchor=(1.05, 1))
plt.tight_layout()
plt.savefig(’average_weeknight_sleep_by_overall_health.png’)



’’’Use a heatmap to visualize the relationship between overall health ratings
and weeknight sleep duration’’’
df_sleep_health = pd.DataFrame(weeknight_sleep_durations, index=health_labels)
df_sleep_health = df_sleep_health.T

plt.figure(figsize=(9,6))
sns.heatmap(
    df_sleep_health,
    cmap="Greens",
    annot=True,
    fmt=".1f",
    cbar_kws={’label’: ’Percentage of Students (%)’}
)

plt.title(’Average Weeknight Sleep Duration by Overall Health Rating’,
          fontsize=14, weight=’bold’)
plt.xlabel(’Overall Health Rating’, fontsize=12)
plt.ylabel(’Average Weeknight Sleep Duration’,fontsize=12)
plt.tight_layout()
plt.savefig(’heatmap_weeknight_sleep_by_overall_health.png’)



’’’Use a stacked bar chart to visualize the relationship between time spent
partying and weekend night sleep duration’’’
partying_per_week_labels = [’0-5 hrs’, ’6-15 hrs’, ’16-25 hrs’, ’26+ hrs’]
weekend_sleep_durations = {
    "<=4h": np.array([0.82, 1.47, 3.34, 6.29]),
    "4h": np.array([1.74, 3.60, 7.69, 8.61]),
    "5h": np.array([5.10, 8.91, 13.04, 14.57]),
    "6h": np.array([12.38, 15.82, 18.28, 14.90]),
    "7h": np.array([21.42, 20.29, 19.73, 17.22]),
    "8h": np.array([30.52, 25.93, 18.17, 17.88]),
    "9h": np.array([19.32, 16.24, 10.81, 7.62]),
    ">9h": np.array([8.70, 7.74, 8.92, 12.91])
}

width = 0.6

fig, ax = plt.subplots(figsize=(10,6))
bottom = np.zeros(len(partying_per_week_labels))

colors = plt.cm.nipy_spectral(np.linspace(0, 1, len(weekend_sleep_durations)))
for i, (timerange, percent) in enumerate(weekend_sleep_durations.items()):
    ax.bar(partying_per_week_labels, percent, width, label=timerange,
           bottom=bottom, color=colors[i])
    bottom += percent

ax.set_title('Average Weekend Sleep Duration by Time Spent Partying Per Week',
             fontsize=13, weight='bold')
ax.set_ylabel('Percentage of Students (%)', fontsize=12)
ax.set_xlabel('Time Spent Partying Per Week', fontsize=12)
ax.set_ylim(0, 100)
ax.legend(title='Average Weekend Night Sleep Duration', loc='upper left',
          bbox_to_anchor=(1.05, 1))
plt.tight_layout()
plt.savefig('average_weekend_sleep_by_partying_per_week.png')



'''Use a stacked bar chart to visualize the relationship between rape trauma
and weekend night sleep duration'''
rape_trauma_labels = ['Not a victim of rape (last 12 months)',
                      'Victims of rape (last 12 months)']
weekend_sleep_durations = {
    "<=4h": np.array([0.88, 3.43]),
    "4h": np.array([1.90, 6.46]),
    "5h": np.array([5.39, 10.51]),
    "6h": np.array([12.65, 16.62]),
    "7h": np.array([21.38, 17.38]),
    "8h": np.array([30.15, 21.63]),
    "9h": np.array([19.02, 15.11]),
    ">9h": np.array([8.64, 8.86])
}

width = 0.6

fig, ax = plt.subplots(figsize=(10,6))
bottom = np.zeros(len(rape_trauma_labels))

colors = plt.cm.nipy_spectral(np.linspace(0, 1, len(weekend_sleep_durations)))
for i, (timerange, percent) in enumerate(weekend_sleep_durations.items()):
    ax.bar(rape_trauma_labels, percent, width, label=timerange, bottom=bottom,
           color=colors[i])
    bottom += percent
```

```
ax.set_title('Average Weekend Sleep Duration by Presence of Rape Trauma',
             fontsize=14, weight='bold')
ax.set_ylabel('Percentage of Students (%)', fontsize=12)
ax.set_xlabel('Presence of Rape Trauma', fontsize=12)
ax.set_ylim(0, 100)
ax.legend(title='Average Weekend Night Sleep Duration', loc='upper left',
            bbox_to_anchor=(1.05, 1))
plt.tight_layout()
plt.savefig('average_weekend_sleep_by_rape_trauma.png')



'''Use a stacked bar chart to visualize the relationship between resilience
level and weeknight sleep duration'''
resilience_labels = ['not resilient', 'rarely resilient',
                     'sometimes resilient', 'often resilient',
                     'very resilient']
weeknight_sleep_durations = {
   "<=4h": np.array([6.03,  2.64,  1.51,  0.87,  1.04]),
    "4h":   np.array([7.07,  5.35,  4.09,  2.77,  2.73]),
    "5h":   np.array([15.49, 14.45, 13.81, 10.87, 10.72]),
    "6h":   np.array([25.61, 29.94, 28.62, 26.41, 25.29]),
    "7h":   np.array([22.26, 26.08, 30.80, 34.62, 33.74]),
    "8h":   np.array([15.41, 14.37, 15.51, 19.24, 21.12]),
    "9h":   np.array([4.54,  4.55,  3.85,  4.01,  4.12]),
    ">9h":  np.array([3.57,  2.64,  1.79,  1.20,  1.24])
}

width = 0.6

fig, ax = plt.subplots(figsize=(10,6))
bottom = np.zeros(len(resilience_labels))

colors = plt.cm.nipy_spectral(np.linspace(0, 1, len(weeknight_sleep_durations)))
for i, (timerange, percent) in enumerate(weeknight_sleep_durations.items()):
    ax.bar(resilience_labels, percent, width, label=timerange, bottom=bottom,
           color=colors[i])
    bottom += percent

ax.set_title('Average Weeknight Sleep Duration by Level of Resilience',
             fontsize=14, weight='bold')
ax.set_ylabel('Percentage of Students (%)', fontsize=12)
ax.set_xlabel('Level of Resilience', fontsize=12)
plt.xticks(rotation=45)
ax.set_ylim(0, 100)
ax.legend(title='Average Weeknight Sleep Duration', loc='upper left',
          bbox_to_anchor=(1.05, 1))
plt.tight_layout()
plt.savefig('average_weeknight_sleep_by_resilience.png')
```

```
'''Use a heatmap to visualize the relationship between resilience level
and weeknight sleep duration'''
df_sleep_health = pd.DataFrame(weeknight_sleep_durations,
                               index=resilience_labels)
df_sleep_health = df_sleep_health.T

plt.figure(figsize=(9,6))
sns.heatmap(
    df_sleep_health,
    cmap="Greens",
    annot=True,
    fmt=".1f",
    cbar_kws={'label': 'Percentage of Students (%)'}
)

plt.title('Average Weeknight Sleep Duration by Level of Resilience',
          fontsize=14, weight='bold')
plt.xlabel('Overall Health Rating', fontsize=12)
plt.ylabel('Average Weeknight Sleep Duration',fontsize=12)
plt.tight_layout()
plt.xticks(rotation=45)
plt.savefig('heatmap_weeknight_sleep_by_resilience.png')



'''Use a stacked bar chart to visualize the relationship between stress
and weeknight sleep duration'''
stress_labels = ['no stress', 'low stress', 'moderate stress', 'high stress']
weeknight_sleep_durations = {
    "<=4h": np.array([3.39,   0.58,  0.69,   2.50]),
    "4h":   np.array([5.41,   1.56,   2.40,   5.92]),
    "5h":   np.array([9.62,   6.29,  10.62, 18.09]),
    "6h":   np.array([21.00,  21.36,  27.17,  30.39]),
    "7h":   np.array([25.21,  37.75,  34.84,  26.21]),
    "8h":   np.array([25.70,  25.95,  18.87,  12.46]),
    "9h":   np.array([6.12,   5.17,   4.08,  2.92]),
    ">9h":  np.array([3.55,   1.35,   1.33,   1.51])
}

width = 0.6

fig, ax = plt.subplots(figsize=(10,6))
bottom = np.zeros(len(stress_labels))

colors = plt.cm.nipy_spectral(np.linspace(0, 1, len(weeknight_sleep_durations)))
for i, (timerange, percent) in enumerate(weeknight_sleep_durations.items()):
    ax.bar(stress_labels, percent, width, label=timerange, bottom=bottom,
           color=colors[i])
    bottom += percent
```

```
ax.set_title('Average Weeknight Sleep Duration by Level of Stress',
             fontsize=14, weight='bold')
ax.set_ylabel('Percentage of Students (%)', fontsize=12)
ax.set_xlabel('Level of Stress', fontsize=12)
ax.set_ylim(0, 100)
ax.legend(title='Average Weeknight Sleep Duration', loc='upper left',
          bbox_to_anchor=(1.05, 1))
plt.tight_layout()
plt.savefig('average_weeknight_sleep_by_stress.png')



'''Use a least-squares polynomial regression to visualize the relationship
between resilience level and weeknight sleep duration'''
diener_scores = np.array([
    8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26,
    27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44,
    45, 46, 47, 48, 49, 50, 51, 52, 53, 54, 55, 56
])
weeknight_sleep_means = np.array([
    2.20, 2.26, 2.00, 1.79, 1.78, 2.05, 2.03, 1.96, 2.05, 1.88, 2.04, 2.09,
    2.12, 2.21, 2.14, 2.05, 2.15, 2.11, 2.19, 2.15, 2.16, 2.14, 2.19, 2.17,
    2.24, 2.25, 2.29, 2.30, 2.29, 2.31, 2.35, 2.38, 2.39, 2.38, 2.41, 2.46,
    2.44, 2.46, 2.50, 2.53, 2.56, 2.56, 2.56, 2.58, 2.57, 2.58, 2.57, 2.64,
    2.61
])

width = 0.6

fig, ax = plt.subplots(figsize=(10,6))
z = np.polyfit(diener_scores, weeknight_sleep_means, 1)
# perform a least-squares polynomial regression fit with the given data points

p = np.poly1d(z)
# create an polynomial function object to be used below

ax.plot(diener_scores, weeknight_sleep_means, color='royalblue', marker='o')
# plot the original data points as a scatter line

ax.plot(diener_scores, p(diener_scores), color='green', linestyle='--',
        label='Trend line')
# plot the trend line produced by the least-squares polynomial regression

ax.set_title('Mean Weeknight Sleep vs. DIENER Score', fontsize=14,
             weight='bold')
ax.set_xlabel('DIENER Score (Life Satisfaction: Higher number = ' \
'higher quality of well-being)', fontsize=12)
ax.set_ylabel('Mean Weeknight Sleep Duration', fontsize=12)
ax.set_ylim(1.5, 2.9)
ax.grid(alpha=0.3)
ax.legend()
plt.tight_layout()
plt.savefig('average_weeknight_sleep_vs_diener_scores.png')



'''Use a barplot to visualize the relationship between students with each
disorder that either also reported not having sleep apnea or did also
report having sleep apnea'''
data = {
    'Disorder': ['ADHD', 'Substance Abuse', 'Anxiety', 'Depression',
                 'Insomnia', 'Academic Impact'],
    'Without_Disorder': [1.86, 2.2, 1.28, 1.26, 1.78, 2.43],
    # % of students with each disorder that did not also report sleep apnea
    'With_Disorder': [5.44, 12.1, 4.38, 5.34, 9.93, 5.66]
    # % of students with each disorder that did also report sleep apnea
}

df = pd.DataFrame(data)

melted_df = df.melt(id_vars='Disorder', value_vars=['Without_Disorder',
            'With_Disorder'], var_name='Group', value_name='Percent')
# melt gives variables two columns each

colors = {'Without_Disorder': 'silver', 'With_Disorder': 'indianred'}

plt.figure(figsize=(9,5))
sns.barplot(x='Disorder', y='Percent', hue='Group', data=melted_df,
            palette=colors)

plt.title('Prevalence of Sleep Apnea by Mental/Behavioral Disorder',
          weight="bold", fontsize="medium")
plt.ylabel('Percent of Students Reporting Sleep Apnea (%)')
plt.xlabel('Diagnosis')
plt.xticks(rotation=30, ha='right')
plt.legend(title='Group')
plt.ylim(0, 35)

plt.tight_layout()
plt.savefig('sleep_apnea_and_other_disorders.png')



'''Use a barplot to visualize the relationship between students with each
disorder that either also reported not having insomnia or did also
report having insomnia'''
data = {
    'Disorder': ['ADHD', 'Substance Abuse', 'Anxiety', 'Depression',
                 'Sleep Apnea', 'Academic Impact'],
    'Without_Disorder': [5.14, 6.75, 1.87, 2.37, 6.6, 5.27],
    # % of students WITHOUT the disorder who report insomnia
    ’With_Disorder’: [19.72, 33.33, 16.95, 20.34, 30.1, 21.64]
    # % of students WITH the disorder who report insomnia
}

df = pd.DataFrame(data)

melted_df = df.melt(id_vars=’Disorder’, value_vars=[’Without_Disorder’,
            ’With_Disorder’], var_name=’Group’, value_name=’Percent’)
# melt gives variables two columns each

colors = {’Without_Disorder’: ’silver’, ’With_Disorder’: ’dodgerblue’}

plt.figure(figsize=(9,5))
sns.barplot(x=’Disorder’, y=’Percent’, hue=’Group’, data=melted_df,
            palette=colors)

plt.title(’Prevalence of Insomnia by Mental/Behavioral Disorder’,
          weight="bold", fontsize="medium")
plt.ylabel(’Percent of Students Reporting Insomnia (%)’)
plt.xlabel(’Diagnosis’)
plt.xticks(rotation=30, ha=’right’)
plt.legend(title=’Group’)
plt.ylim(0, 35)

plt.tight_layout()
plt.savefig(’insomnia_and_other_disorders.png’)

’’’
AI assistance was used to suggest solutions and resolve errors
’’’
```

7.4 Appendix D: Linear Regression Modeling

```
'''
Jonah Watson
Fall 2025
Sleep Health in College Students: A Multivariable Predictive Modeling Analysis
This file was used to create and test the linear regression model used in this
research project.
'''

import pandas as pd
# imported for data manipulation and analysis

from sklearn.model_selection import train_test_split
# imported to efficiently split the data into training and testing groups

from sklearn.linear_model import LinearRegression
# imported to use the linear regression model

from sklearn.metrics import mean_absolute_error
 # imported to calculate the mean absolute error score


# load the cleaned dataset
df = pd.read_csv("CLEANED SLEEP NCHA-III S24 - New_Numeric.csv")



'''Create predictor groups for the linear regression model'''
psych = ['N3Q42B', 'N3Q48', 'DIENER', 'RKESSLER6', 'RULS3']
# resilience, stress, well-being/life satisfaction, serious psychological...
# ...distress, and loneliness

behavior = ['N3Q1', 'N3Q3E', 'N3Q3I', 'N3Q13_recode']
# overall self-rated health, time spent doing physical activity, time spent...
# ...partying, and time taken to fall asleep

diagnoses = ['N3Q65A2', 'N3Q65A3', 'N3Q65A7', 'N3Q65A15','N3Q65A28',
            'N3Q65A35']
# ADHD, substance abuse, anxiety, depression, insomnia, and sleep apnea

trauma = ['N3Q20D', 'N3Q20F', 'N3Q20G']
# sexual assault, rape, and stalking victims



'''Write predictor groups into a dictionary to properly iterate'''
groups = {
    "Psychological" : psych,
    "Behavioral" : behavior,
    # "Diagnoses" : diagnoses,
    "Trauma" : trauma
}

# define the target of the prediction (y)
y = df[’N3Q14’]     # weeknight sleep duration

# cumulative groups
all_predictors = []
results = []

for name, group in groups.items():
    all_predictors.extend(group)
    # with each loop iteration, attach the next group that was defined above
    X = df[all_predictors]

    # drop missing predictor and/or target values
    print("Original data size:", len(X))
    df_model = pd.concat([X, y], axis=1).dropna()
    X = df_model[X.columns]
    y_clean = df_model[y.name]
    print("After dropping NaNs:", len(X))

    # split the data into test and train groups: randomly divide the data...
    # ...into 80% used for training and 20% used for testing
    X_train, X_test, y_train, y_test = train_test_split(X, y_clean,
                                        test_size = 0.2, random_state=67)

    # train the regression model
    model = LinearRegression()
    model.fit(X_train, y_train)
    # model.fit finds the best-fitting coefficients to minimize error in training

    # save coefficients for each model (print slope for current predictor group)
    coefficient_table = pd.DataFrame({
        ’Feature’: X.columns,
        ’Coefficient’: model.coef_
    })

    print(f"\n--- Coefficients for {group} ---")
    print(coefficient_table)

    # evaluate
    y_prediction = model.predict(X_test)
    r2 = model.score(X_test, y_test)
    n = X_test.shape[0]   # number of samples
    p = X_test.shape[1]   # number of predictors
    adjusted_r2 = 1 - (1 - r2) * ((n - 1) / (n - p - 1))
    # r-squared measures variance, where 1 is perfect, 0 is as if you guessed...
    # ...randomly, and -1 is worse than guessing the mean
    # "adjusting" r-squared according to this formula allows it to more
    # effectively display how certain variables are contributing to the model
    MAE = mean_absolute_error(y_test, y_prediction)
    # MAE measures how far off predictions are, on average...
    # ...from what the true values are (lower = better)

    results.append({
        "Group Added": name,
        "Number of Predictors": len(all_predictors),
        "Adjusted R²": adjusted_r2,
        "Mean Absolute Error": MAE,
        "Coefficients": coefficient_table
    })

print("\n================")
print("Summary of New Results by Group")
print("================\n")

"""The following for loop shows how much each predictor contributes to the
prediction. For example if stress is -0.21 and resilience is +0.09 it means
that higher stress decreases predicted sleep by -0.21 of a scale point,
which, in this case,where 1 hour more of sleep is 1 scale point means that
-0.21 is about 12 minutes less of sleep."""
for result in results:
    print(f"\nResults after {result[’Group Added’]}:\n")
    print(f"Adjusted R²: {result[’Adjusted R²’]}")
    print(f"Mean Absolute Error = {result[’Mean Absolute Error’]}")
    print(result[’Coefficients’])
    print("\n")

’’’
AI assistance was used to suggest solutions and resolve errors
’’’
```

7.5 Appendix E: Logistic Regression Modeling

```
'''
Jonah Watson
Fall 2025
Sleep Health in College Students: A Multivariable Predictive Modeling Analysis
This file was the second iteration of modeling practices, this time testing
a logistic regression model, lasso, and decision tree.
'''

from sklearn.metrics import accuracy_score, classification_report, confusion_matrix
from sklearn.metrics import ConfusionMatrixDisplay, roc_auc_score, RocCurveDisplay
# imported for evaluation metrics and access to the confusion matrix and its display

from sklearn.tree import DecisionTreeClassifier
# imported for integration of decision tree

from sklearn.linear_model import LogisticRegression
# imported for logistic regression modeling

from sklearn.pipeline import make_pipeline
# imported for creating a machine learning pipeline

from sklearn.impute import SimpleImputer
# imported for handling missing data

from sklearn.preprocessing import StandardScaler
# imported for feature scaling

from sklearn.model_selection import train_test_split
# imported for splitting data into training and testing sets

import matplotlib.pyplot as plt
# imported for visualizing confusion matrix

import numpy as np
# imported for numerical operations

import pandas as pd
# imported for data manipulation

import statsmodels.api as sm
# imported for statistical modeling



# load the cleaned dataset
df = pd.read_csv("CLEANED SLEEP NCHA-III S24 - New_Numeric.csv")
```

```
# choose which variables will be included in the modeling process
df_filtered = df[[’N3Q1’, ’N3Q14’, ’N3Q42B’, ’N3Q48’, ’N3Q20D’, ’N3Q20F’,
                ’N3Q20G’, ’N3Q65A2’, ’N3Q65A3’, ’N3Q65A7’, ’N3Q65A15’,
                ’N3Q65A28’, ’N3Q65A35’, ’RULS3’]]

"""Recode the target variable (Weeknight sleep duration) as this model does
not include students receiving greater than 9 hours of sleep per night"""
df_filtered = df[df[’N3Q14’] != 8]



print(df_filtered[’N3Q14’].value_counts())
df_filtered[’sleep_binary’] = df_filtered[’N3Q14’].apply(lambda x: 1 if x >= 5 else 0)
# lambda function to create binary target variable
# 1 represents if a student averages 7-9 hours of sleep per night (optimal)
# 0 represents if a student obtains 6 hours of sleep or less per night (suboptimal)
print(df_filtered[’sleep_binary’].value_counts())

"""Recode binary variables so that their responses switch from 1s and 2s to
0s and 1s, where 1 will indicate the presence of the condition"""
binary_vars = [’N3Q20F’, ’N3Q20G’, ’N3Q65A2’, ’N3Q65A3’,
               ’N3Q65A7’, ’N3Q65A15’, ’N3Q65A28’, ’N3Q65A35’, ’RULS3’]
for col in binary_vars:
    df_filtered[col] = df_filtered[col].replace({1: 0, 2: 1})

"""Recode this question so that higher values indicate better overall health
to be consistent for the format of other questions"""
df_filtered[’N3Q1’] = 6 - df_filtered[’N3Q1’]



"""The next few lines of code separate the predictors (X) and the target
variable (y) from each other. X contains features that are related to
psychological, behavioral, diagnostic, and trauma-related factors and y is
the binary indicator for whether the student align with optimal sleep
duration or not."""
X = df_filtered[[’N3Q1’, ’N3Q42B’, ’N3Q48’, ’N3Q20D’, ’N3Q20F’, ’N3Q20G’,
                 ’N3Q65A2’, ’N3Q65A3’, ’N3Q65A7’, ’N3Q65A15’, ’N3Q65A28’,
                 ’N3Q65A35’, ’RULS3’]]
y = df_filtered[’sleep_binary’]

"""Split the data into training and test sets, stratify=y balances
train/test proportions"""
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2,
                                                stratify=y, random_state=67)

"""The following line of code creates a pipeline for the model. SimpleImputer
handles missing values by replacing them with the mean of the column values.
StandardScaler standardizes features to make scaling consistent.
LogisticRegression is the model being used here to predict the probability of
being in the optimal sleep category, using a max of 1000 iterations."""
```

```
model = make_pipeline(SimpleImputer(strategy='mean'), StandardScaler(),
                      LogisticRegression(max_iter=1000))
model.fit(X_train, y_train)

y_pred = model.predict(X_test)
# prediction of class 0 or class 1 as defined above
y_proba = model.predict_proba(X_test)[:, 1]
# probability of being in optimal sleep range for each test case (1)

# ===================
# EVALUATION METRICS
# ===================

print("Accuracy: ", accuracy_score(y_test, y_pred))
# proportion of total correct predictions

print("ROC-AUC: ", roc_auc_score(y_test, y_proba))
# how well the model separates the two classes

print(confusion_matrix(y_test, y_pred))
print(classification_report(y_test, y_pred))

# ROC-AUC curve
RocCurveDisplay.from_predictions(y_test, y_proba)
plt.title("ROC Curve for Logistic Regression Sleep Model")
plt.show()



# extract the logistic regression step from the pipeline
log_reg = model.named_steps['logisticregression']

"""See how much each feature contributes to the prediction,
coefficients > 0 increase the odds of optimal sleep and vice versa"""
coefs = pd.DataFrame({
    'Feature': X.columns,
    'Coefficient': log_reg.coef_[0],
})

"""Convert data to odds ratios for easier interpretation, where an odds
ratio > 1 means that the predictor increases probability of optimal sleep"""
coefs['Odds_Ratio'] = np.exp(coefs['Coefficient'])
coefs = coefs.sort_values(by='Odds_Ratio', ascending=False)
print(coefs)

# compute a confusion matrix for easier interpretation
cm = confusion_matrix(y_test, y_pred)
print(f"Confusion matrix: {cm}")

# visualize the confusion matrix and provide clearer labels
cmdisplay = ConfusionMatrixDisplay(confusion_matrix=cm,
            display_labels=['Suboptimal Sleep (0)', 'Optimal Sleep (1)'])
cmdisplay.plot(cmap='Blues', values_format='d')
plt.title('Confusion Matrix for Logistic Regression Sleep Model')
plt.show()



"""Below is the decision tree classification pipeline. Like the logistic
regression model, missing values will be replaced with the column mean using
SimpleImputer(strategy='mean'). Gini is used to determine the purity of the
data. Having no max_depth value of None allows the tree to grow fully."""
tree = make_pipeline(SimpleImputer(strategy='mean'), DecisionTreeClassifier(
    criterion='gini', max_depth=None, min_samples_split=2, random_state=67))

tree.fit(X_train, y_train)

# calculate predictions and performance metrics for the tree model
y_pred = tree.predict(X_test)
print(classification_report(y_test, y_pred))

# visualize the confusion matrix
cm = confusion_matrix(y_test, y_pred)
display = ConfusionMatrixDisplay(cm, display_labels=['Suboptimal (0)',
                    'Optimal (1)']).plot(cmap='Purples', values_format='d')
plt.title("Decision Tree Confusion Matrix")
plt.show()



# random forest implementation
print("\nRANDOM FOREST IMPLEMENTATION\n")

"""Below is the random forest pipeline. SimpleImputer(strategy='mean') will again
fill missing values with the column mean. RandomForestClassifier creates a model
utilizing many decision trees."""
rf_model = make_pipeline(
    SimpleImputer(strategy='mean'),
    RandomForestClassifier(
        n_estimators=300,          # number of trees
        max_depth=None,            # allow full tree growth
        min_samples_split=2,       # default split rule
        random_state=67,           # for reproducibility
        class_weight="balanced"    # handles imbalance in sleep_binary
    )
)

# train the model
rf_model.fit(X_train, y_train)

# predictions
rf_pred = rf_model.predict(X_test)
```

```
rf_proba = rf_model.predict_proba(X_test)[:, 1]

# evaluation metrics
print("Accuracy:", accuracy_score(y_test, rf_pred))
print("ROC-AUC:", roc_auc_score(y_test, rf_proba))
print(classification_report(y_test, rf_pred))

# confusion matrix
rf_cm = confusion_matrix(y_test, rf_pred)
display = ConfusionMatrixDisplay(rf_cm, display_labels=[’Suboptimal (0)’,
                                                        ’Optimal (1)’])
display.plot(cmap=’Greens’, values_format=’d’)
plt.title("Random Forest Confusion Matrix")
plt.show()


# FEATURE IMPORTANCE PLOT

# extract the trained random forest object
rf = rf_model.named_steps[’randomforestclassifier’]

importances = rf.feature_importances_
feature_names = X.columns

# package into a DataFrame
rf_imp = pd.DataFrame({
    ’Feature’: feature_names,
    ’Importance’: importances
}).sort_values(by=’Importance’, ascending=True)

# horizontal bar chart
plt.figure(figsize=(10, 6))
plt.barh(rf_imp[’Feature’], rf_imp[’Importance’])
plt.xlabel("Feature Importance Score")
plt.title("Random Forest Feature Importance")
plt.show()



"""Shown below is the attempt at integrating LASSO with the logistic
regression model, which did not modify results at all and is mostly
omitted from the report."""

’’’
lasso_model = make_pipeline(SimpleImputer(strategy=’mean’), StandardScaler(),
LogisticRegression(penalty=’l1’, solver=’liblinear’, C=1.0,
max_iter=1000, random_state=67))

lasso_model.fit(X_train, y_train)
y_pred = lasso_model.predict(X_test)
```

```
print(classification_report(y_test, y_pred))
cm = confusion_matrix(y_test, y_pred)
ConfusionMatrixDisplay(cm, display_labels=['Suboptimal (0)',
'Optimal (1)']).plot(cmap='Purples', values_format='d')
plt.show()
'''

'''
AI assistance was used to suggest solutions and resolve errors
'''
```