Investigating and Predicting Alcohol Misuse Risk Among American College Students

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

Alcohol consumption is prevalent among American college students and is associated with academic, behavioral, and mental health consequences. Prior research identifies connections between alcohol misuse and factors such as stress, mental illness, and campus environments. This study investigates the mental, behavioral, and academic factors associated with alcohol consumption and risk of alcohol misuse among American college students. Data were utilized from the American College Health Association's National College Health Assessment III (ACHA–NCHA III), Spring 2024 survey. Alcohol-related behaviors were examined using measures of drinking frequency, binge drinking, and a composite risk score for alcohol misuse. Psychological, diagnostic, behavioral, academic and environmental predictors were incorporated. An ordinal logistic regression model was developed to classify students as low, moderate, or high risk for alcohol misuse. The model demonstrated strong ordinal predictive power, with a MAE score of 0.278 and a within-1 category proportion of 98.7%, although instability was observed when predicting higher-risk students. Suicidal behavior and involvement in Greek life increased the odds of greater alcohol risk by 20%, representing the strongest predictors. Psychological distress (+10%), loneliness (+9%), lower GPA (+8%), lower well-being (+8%), and depression (+7%) were also associated with elevated alcohol risk. Poor mental health symptoms were stronger predictors than mental health diagnoses themselves. PTSD, insomnia, sleep apnea, OCD, and resilience scores showed little to no predictive power. These findings reinforce prior literature, while also opening doors for future research, highlighting the importance of data-driven campus interventions addressing alcohol-related behaviors and underlying mental health factors.

...

Table of Contents

1 Introduction

Alcohol has been a long-standing drug in American culture, and has been even longer standing in many other cultures around the world. Alcohol is consumed legally by millions of people around the world and is often considered a social drug. Alcohol, however, comes with a majority of risks. Alcohol affects the human brain in many ways. Alcohol is a central nervous system depressant, meaning that it slows brain activity, and in turn can change mood, behavior, and self-control. It reduces overall brain function, impairing thinking and reasoning as well as muscle coordination and physical control. The nuanced effects of alcohol vary from person to person based on frequency of drinking, number of drinks, age, sex, genetics, whether a family history of drinking problems exists, and overall health. A standard drink in the United States contains 14 grams of pure alcohol, found in 12 ounces of beer (5% alcohol content), 5 ounces of wine (12% alcohol content) or 1.5 ounces of a shot of liquor or distilled spirits (40% alcohol content). Moderate drinking is considered to be no more than 2 standard drinks a day for men or 1 standard drink a day for women. Binge drinking, a form of excessive drinking, is considered to occur when someone has a Blood Alcohol Concentration (BAC) of 0.08% or greater, which typically occurs after 5 or more drinks within a few hours for a man, or 4 or more drinks within a few hours for a woman. Binge drinking increases the risk of injuries, violent behavior, vehicle accidents, and overdose. Driving at this BAC or higher is illegal in all states in the United States and is considered to be the point where vehicle accident risk increases exponentially [1]. In 2023, 12,429 people died from drunk driving incidents in the United States alone, which averages one death by drunk driving every 42 minutes. There 2,117 alcohol-related deaths where the driver at fault had a BAC between 0.01 and 0.07 [2]. All of these deaths could have been preventable. Heavy alcohol use counts as 5 or more drinks on any day or more than 15 drinks per week for men, or 4 or more drinks on any day or more than 8 drinks per week for women, and over a long period of time can lead to alcohol use disorder (AUD), an increased risk of injuries and certain cancers, liver and heart disease, and other health issues. People who should not drink alcohol include those with compulsions to drink or inability to control the amount they drink, those with medical conditions that may worsen from alcohol consumption, those taking medications that may interact negatively with alcohol, those in recovery from AUD, those who plan on driving or operating machinery, those under 21 years of age, and women who are pregnant. [1].

Due to all of the aforementioned issues, alcohol is a public health concern. For college students, alcohol use is commonplace, and many students use it to socialize, especially on weekends, holidays, and before school events such as homecoming. 4.4 million college students aged 18 to 25 (49.6%) reported drinking alcohol at least once in a given month, with 2.6 million college students (29.3%) reported at least one case of binge drinking. Additionally, 608,000 college students aged 18 to 25 (6.8%) reported heavy drinking at least once in a given month, and 14.5% of full-time college students meet the criteria for past-year alcohol use disorder (AUD). Furthermore, although reporting numbers on sexual assault is very challenging due to the frequency of under-reported cases, it has been confirmed that one in five women experiences sexual assault in college, with the majority of these cases involving alcohol or other substances [3]. In addition, other negative consequences that may arise from student alcohol consumption habits include academic issues such as missing classes and receiving poor grades [4], various mental

health issues [5] and, in some cases, more specifically depressive issues [6], and other issues such as browning out, making regretful choices, having unprotected sex, blacking out, physically injuring themselves [7], and causing property damage [8]. In terms of contributors to alcohol related issues in college students, living arrangements, college environments and longer exposure to them, and peer pressure are common contributors [9]. These college environments may include participation in athletics, Greek Life, and hazing rituals. Genetic predisposition and use in high school can also contribute to alcohol misuse in college [4], as can parental alcohol use, mental health problems [5] and stress [6].

Although alcohol consumption and misuse rates among college students may not be as high as they used to be, the issue of alcohol misuse in college persists. Previous studies have covered a multitude of contributors to alcohol misuse and consequences of alcohol misuse, but as the issue persists, studies should continue to be conducted to evaluate the best methods to remedy these issues. The purpose of this study is to provide new information on the relationship between college students and their alcohol consumption, but also to add further support to previous studies, ultimately pushing research forward and also allowing students who struggle with alcohol misuse to receive proper support and overcome their difficulties. This study will begin with a review of relevant literature and then move to the methodology and exploratory data analysis (EDA) and visualization sections. From the findings of the EDA, a model will be created and tested to identify relevant insights and the impact of variables. The choice between a classification model and a predictive model will be decided after analysis and visualization are conducted. From there, results will be discussed and connections will be made to prior research, followed by concluding remarks.

In the next section, relevant studies on this topic will be discussed in more detail.

2 Literature Review

This section serves as a review of previous studies that involve alcohol misuse, especially among college student populations, covering both the contributors to alcohol misuse and also the consequences of alcohol misuse. Many specific factors will be discussed in this literature review, and a large number of these factors will be investigated in the data analysis and visualization section. These include, but are not limited to, mental, behavioral, psychological, physical, and environmental factors.

The study by Lorant et al., [9], explored the impact of environmental factors on student drinking behavior. An online questionnaire was distributed to all students at a university in Belgium and 7,015 students participated. The survey included issues related to drinking behavior, social involvement, environmental factors, drinking norms, and perceived consequences of drinking. This research revealed that students consumed an average of 1.7 alcoholic drinks per day and engaged in abusive drinking 2.8 times per month. Drinking behavior was significantly impacted by living arrangements, where students who lived in dormitories or with more roommates were more likely to drink frequently and heavily. The prevalence of heavy drinking also increased with longer exposure to the college environment (more years enrolled). Celebratory rituals and hazing were associated with increased alcohol use (referred to as "traditional student folklore"). The authors of the

article advocate for an increased responsibility for universities to shape the environment to reduce social and environmental pressures to drink, and that the alcohol consumption habits of college students cannot be solely framed as a personal choice [9].

The study by White and Hingson, [4], examined the burdens and consequences commonly associated with student alcohol use through a comprehensive literature review. There were several key findings from this research. First, the drinking behavior of college students is influenced by several factors including but not limited to genetic predisposition, use during high school, and college environmental aspects like campus norms, participation in Greek life and athletics, and the availability of alcohol. There are many consequences that can arise from the excessive drinking habits of college students. There may be academic issues like missing classes and receiving poor grades, health issues such as injuries, overdose, and long-term cognitive deficits, as well as social issues like cases of sexual assault or fatalities. The authors advocate for further research into improved methodologies and methods to combat alcohol misuse within college communities. Further research would allow students and communities to more accurately assess and address alcohol misuse within college communities. This requires a collaborative effort between colleges and the community, made possible by both policy changes and targeted initiatives [4].

The study by Herrero-Montes et al., [10], studied the excessive alcohol use and binge drinking among students. The study used the Alcohol Use Disorders Identification Test (AUDIT), an official questionnaire from the National Institute on Drug Abuse (NIDA). The study organizers surveyed 142 students, with the majority being women, and made a couple key findings, these being that 38% of participants participated in binge drinking, with a significant proportion of these participants also meeting the criteria for harmful alcohol use or alcohol use disorder. From the AUDIT questionnaire results of the binge drinkers, the data suggested that there were deeper patterns of problematic alcohol use in this subset. The authors discuss the vulnerability of healthcare students to risky drinking behaviors, with long-term harmful implications on both individual health and future professional roles. They also advocate for educational and preventative interventions inside of campus communities to prevent unhealthy drinking habits from becoming habitual in students, especially for those who are more vulnerable to developing dependencies or patterns of misuse [10].

The study by Davoren et al., [11], reviewed alcohol consumption habits among university students in Ireland and the United Kingdom between 2002 and 2014. The study pulls from 29 cross-sectional studies identified through major databases such as MEDLINE, EMBASE, CINAHL, and PsychInfo. The author's intentions were to assess the prevalence of hazardous alcohol use in this population and increase awareness around health policies and interventions. The studies utilized in this report used various validated screening tools, such as the Alcohol Use Disorders Identification Test (AUDIT), the CAGE questionnaire, and measures of weekly drinking limits. The findings revealed that around two-thirds of university students were classified as hazardous drinkers according to the criteria of the AUDIT, with the CAGE questionnaire suggesting that over 20% of the students have experienced alcohol problems for several years. Another key finding was that more than 20% of students exceeded realistic weekly drinking limits, with a notable trend in the narrowing of the gender gap for hazardous drinking behaviors over the review period. The authors speak urgently about the need for standardized methods in

alcohol research in universities, public health programs, and interventions to help reduce alcohol-related harm among students [11].

The study by Ay et al., [5], investigated the prevalence of hazardous alcohol consumption (HAC) and the factors associated among university students in Turkey. The study was conducted at Eskisehir Osmangazi University during the 2019–2020 academic year, and it utilizes a stratified sampling method to represent 26,036 undergraduate students across 11 disciplines. There were 2,349 students who participated in the survey, which utilized the Alcohol Use Disorders Identification Test (AUDIT) to assess alcohol consumption patterns. Hazardous alcohol consumption (HAC) was defined as an AUDIT score of 8 or more. The findings revealed that 13.5% of the participants met the criteria for HAC, with a higher prevalence among males (18.8%) compared to females (8.2%). The researchers utilized multivariate logistic regression analyses which allowed them to identify several factors associated with an increased risk of HAC in both genders. These factors were the alcohol consumption of parents and close friends, smoking, illegal substance use, and mental health problems. This study shows the need for targeted interventions to mitigate hazardous drinking behaviors among university students in Turkey [5].

The study by Chow et al., [6], examined the connection between alcohol consumption and depressive symptoms among 345 full-time undergraduate students enrolled at The University of Hong Kong. These students completed questionnaires assessing their drinking behavior (AUDIT and CAGE), depressive symptoms (PHQ-9), and stress-coping strategies (COPE inventory). The study also evaluated students' knowledge and perceptions of the health impacts of alcohol. Key findings revealed that 43.2% of students were moderate- to high-risk drinkers, and 57.9% reported mild to moderately severe depressive symptoms. Higher depression scores were linked with students in non-medical disciplines, students who were smokers, those with high general stress levels, and those who used avoidance as a primary stress-coping method. Students commonly underestimated their own drinking behaviors, and avoidance coping was significantly more prevalent in high-risk drinkers. The authors advocate for mental health interventions and improved alcohol education within campus environments, with support systems built to reduce stigma and normalize help-seeking behavior [6].

The study by Kenney et al., [12], investigated connections between the mental health, drinking norms of peers and friends, and individual alcohol consumption of 1,254 first-year college students. The study focused on how students' perceptions of their peers' drinking behaviors influenced their own alcohol use, especially for those experiencing poor mental health. Students in the study were more susceptible to heavy drinking if their friends also engaged in heavy alcohol consumption. Those who had anxiety or depression were especially susceptible to heavy drinking, suggesting that alcohol risk among students is often exacerbated by the drinking norms of their friends, especially for those with poor mental health. This study shows the importance of addressing both mental health and social perceptions of alcohol consumption to reduce risky drinking behaviors on college campuses [12].

Pearson et al., [13] conducted a large-scale study to understand how much alcohol consumption can actually explain the negative consequences students experience, such as hangovers, missed classes, injuries, and risky behaviors. The authors analyzed data from many college student samples with the goal of understanding how different ways of mea-

suring drinking, such as how often someone drinks or how much they drink in one night, relate to harmful outcomes. The results of the study showed that students who drank more often or in larger quantities experienced more alcohol-related problems. Students who reported frequently getting drunk or binge drinking had up to a 47% stronger link to negative consequences compared to those who drank less often. Even with these connections, however, the study found that only about 15% to 23% of the problems students faced could be explained by how much they drank, meaning that the majority (around 77% to 85%) of the negative consequences came from other factors, such as personal habits, mental health, and environmental influences. The authors concluded that simply measuring how much students drink is not enough to fully predict who will experience alcohol-related harm. The factors separate from this need to be considered in programs aimed at reducing alcohol problems to come to a more well-rounded conclusion [13].

Shorter et al., [14], conducted a regional study in County Donegal, Ireland, to explore alcohol consumption patterns and public attitudes toward evidence-based alcohol policies. The goal of the research was to provide localized insight to support the implementation of Ireland's Public Health (Alcohol) Act 2018 by comparing responses from both a sample of students and a sample of adults of the general population. The study showed that hazardous drinking is highly prevalent. 59% of students and 53% of adults scored a 5 or higher on the AUDIT-C, which indicates risky levels of alcohol use. 46% of students and 36% of adults self-reported binge drinking, which is defined as consuming six or more drinks on a single occasion at least once per month. Participants also reported experiencing the secondary effects of others' drinking, where 23–25% of participants stated that someone in their household was a heavy drinker, and 19% reported harm caused by others' alcohol use. Most participants supported public health efforts to mitigate alcohol-related harm, with 58% of students and 53% of adults agreeing that health authorities should intervene. However, only 36% of students and 32% of adults supported minimum unit pricing, a policy in place to help combat alcohol-related harm. The study also found that individuals who had experienced harm due to others' drinking were more likely to support alcohol regulation policies. The authors speak to the need for community-driven and evidence-based public health strategies. They argue that effective alcohol interventions must be rooted in local data and supported by the community to be able to properly address the cultural and social contexts where drinking occurs [14].

The National College Health Assessment (NCHA) III 2022 Executive Summary, [7], summarizes the key findings that come from the survey results, including hazardous activity involving student alcohol consumption, interviewing over 24,000 students from more than 130 collegiate institutions. Key findings in this subject matter include that within 30 days of taking the survey, 17% of students who drank alcohol also drove while under the influence, as well as that survey respondents who drank alcohol in the last year had experienced other issues such as browning out (13,5%), doing something they later regretted (10%), having unprotected sex (6%), blacking out (4.9%), and physically injuring themselves (1.3%). The agenda of the American College Health Association (ACHA) is to collect student health data to assist those in positions of influence in campus communities such as health service providers, educators, administrators, counselors, and wellness consultants [7].

Lukács et al., [15], conducted a study in 2013 to examine alcohol consumption patterns at the University of Miskolc in Hungary. The study includes 658 students with an average

age of 20.6 years old. The survey revealed that over 90% of students drank alcohol with a nearly equal amount having been drunk before. Within the last year, 80% of male students and 64% of female students had reported being drunk more than 10 times. Binge drinking was practiced regularly by 78% of males (five or more drinks on one occasion) and 51% of females (four or more drinks on one occasion), and most students reported that they only drink when with other people, which indicates the prominent social nature of drinking habits. The study also revealed that the average age of first intoxication was 16 years old, several years before university. From this study, the authors concluded that high levels of alcohol consumption are embedded within the campus culture and require targeted interventions and prevention programs to disrupt [15].

The study conducted by Tang et al., [16], in 2023 analyzed survey data from 3,719 college students aged 18-25. The data was captured during the COVID-19 pandemic through the ACHA–NCHA III. This study explored how alcohol consumption and marijuana consumption behaviors intersect with risky driving and substance use. The researchers found that students who engaged in binge drinking were significantly more likely to also report driving after using marijuana, and also found that alcohol use is not always an isolated behavior, and many times is part of a larger pattern of poly-substance use which correlates with risky behaviors including impaired driving. The authors push for prevention messaging that targets marijuana and alcohol use, especially regarding impaired driving [16].

The NIAAA, [17], reported that the majority of U.S. college students, (ages 18–22, consumed alcohol monthly, with 37% of them engaging in binge drinking and roughly 9% of full-time students meeting the criteria for Alcohol Use Disorder (AUD). These behaviors are linked to serious consequences, including approximately 1,500 annual student deaths, 696,000 assaults, and 97,000 cases of sexual assault. Student academic performance is also affected, with binge drinkers noticeably more likely to skip class and perform poorly. Unstructured time, Greek life, athletic culture, and weak enforcement of underage laws are the primary contributors to these outcomes, with the first six weeks of college being especially high-risk for new students. The report emphasizes targeted intervention and prevention strategies to address these risks and minimize alcohol-related harm. [17].

A new NIAAA report was issued in 2023, [18], as part of an annual back-to-college public health campaign. Key findings from this report include that about 60% of college students, aged 18-22, consumed alcohol in the past month with roughly 40% of students engaging in binge drinking (defined as 5+ drinks for males or 4+ for females in one sitting), an increase of roughly 3% from the 2020 report, and also that an estimated 15% of college students met the criteria for Alcohol Use Disorder (AUD), up from the 9% estimate from previous years. Numbers of annual student deaths, assaults, and cases of sexual assault remained showed little to no increase but remain equally pertinent. Alcohol use was shown to correlate with poor academic performance, with many students missing classes and therefore falling behind in their classes as well receiving lower grades. This report was aimed at parents and caregivers and pushes for them to talk with their students before and during the school year about the possible consequences of risky drinking and campus support resources. Parental involvement can be a protective factor, especially for first-year students adjusting to their new environment [18]. Furthermore, in 2025 the NIAAA presented additional and newer findings. Key points include that, in a given month, 4.4 million college students aged 18 to 25 (49.6%) reported drinking

alcohol at least once, 2.6 million college students (29.3%) reported at least one case of binge drinking, and 608,000 college students (6.8%) reported heavy drinking at least once. Additionally, 14.5% of full-time college students meet the criteria for AUD in the past-year. Also reported was that one in five women experiences sexual assault in college, with the majority of these cases involving alcohol or other substances [3]. Even with some slight improvements from the 2023 report, these issues remain just as relevant and important to continue to address in society and on campuses.

In 2016, Yoder et. al, [19], conducted a study on 24 social drinkers and 21 non-treatment-seeking alcoholics (NTS) using Positron Emission Tomography (PET) imaging. PET imaging is used in the medical field to measure changes in metabolic activity inside the body, and was used during intravenous alcohol infusion in this study, where alcohol was administered via IV. The researchers found that NTS alcoholics, who are typically early-phase alcoholics, have hyper-reactive reward systems, which indicate a snowball effect of increasing pleasure aligned with increased drinking, driving compulsive use. Eventually, this pleasure experiences somewhat of a peak and alcohol consumption becomes less of a reward-driven (ventral) behavior and becomes more of a habit-driven (dorsal) behavior, which is where alcoholism becomes harder to treat. Social drinking, widely considered such a large and consistent part of the "college experience", can transition college students from casual drinking to disordered drinking over time [19].

Conroy and de Visser, [20], performed a study in 2017 that analyzed 511 British university students. Around 45% of these students had temporarily abstained from drinking in social situations, and these students reported benefits such as improved physical and psychological health, higher self-esteem, and a more productive and stable social life. Among the women who abstained, there was a notable correlation with their intent to follow low-risk drinking guidelines. This study indicates that choosing to abstain from drinking can improve physical, mental, and behavioral health even after just a week, the time frame in which this study was conducted [20].

De Visser et. al., [21], conducted a study with 4,232 adults who participated in "Dry January" in 2016, a campaign based around abstaining from alcohol in the month of January. The participants completed surveys at the start of the month, the end of the month, and then also six months later. Some of the key findings here were that, across the board, short-term abstinence tended to lead to an improved sense of well-being and self-efficacy, and also that 70% of participants had maintained reduced alcohol consumption up to the six month mark, even if these participants did not fully abstain. Those who successfully abstained during January showed significant drops in their AUDIT scores and had fewer drinking days with less frequent intoxication. This study shows that a short abstinence period can greatly improve or even completely reset unhealthy drinking habits and sustain the improved behavior in the future, with mental health benefits in mind [21].

The following section of the report will cover the methodology used with data collected for firsthand analysis and modeling. This section will cover the step-by-step process used with the data, rationalize choices in regards to included variables, and explain many other elements of this research.

3 Methodology

This study serves as an exploratory data analysis and visualization of raw survey data, which will be followed by a process of model building and testing. A multi-stage data science approach was used to examine alcohol and its relationship to other behaviors and traits in students. The survey data used in this study was purchased from and belongs to the American College Health Association (ACHA). The ACHA administers a National College Health Assessment (NCHA) every spring and fall to American college students. This section of the report will cover the steps of the methodology used in this project. The methodology consisted of many stages, beginning with data collection and profiling to assess the scope and structure of the data, followed by data cleaning and preprocessing to ensure that variables were prepared and fit for analysis. This included the removal of duplicate entries, handling of missing values, and conversion of categorical responses to numerical codes. EDA and visualization were then 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 topics related to alcohol and other topics related to psychology, behavior, and mental health. Finally, the data was polished and refined, so that the significant findings and implications on college health research can be presented with clarity. See Figure 1 for the outline description used to work through the data used in this study.

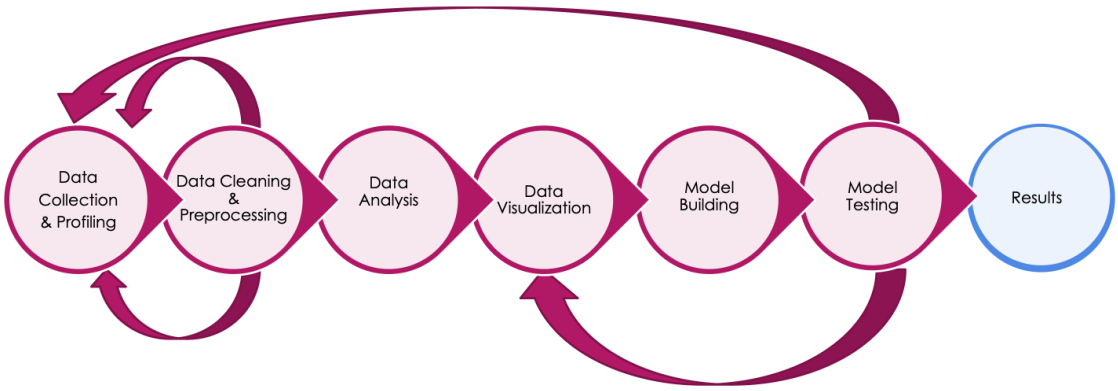

Figure 1: The process used in tandem with the data for this study.

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, where the survey was administered between January and June. This survey includes responses from 103,639 students representing 154 American higher education institutions, making it one of the largest NCHA samples to date. 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 Python pandas module, 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 initial group of variables that were received was reduced in size as the study continued and the focus narrowed.

Tables 1-3 contain response rates for all the variables of interest for this study from the ACHA-NCHA spring 2024 survey. Table 1 displays the distribution of general demographics, Table 2 displays the distribution of alcohol-related behaviors among participants, and Table 3 displays the distribution of responses of questions related mental health and general health. Note that percentages for each question often do not add up to 100% as missing values, which did not feel necessary to show, make up the rest of that 100%. Missing values will be dealt with on a case-by-case basis so that if one column has missing data, the rest of the row can still contribute to other parts of the analysis process. Refer to Appendix B: Profiling for the Python code used for this section of the pipeline. For additional details on these survey questions, 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: Demographics

Study Sample Size	103,639
N3Q67A: What sex? — female 1, male 2	
Female	72,510 (69.9%)
Male	29,753 (28.7%)
N3Q69: How old are you? — (Numeric value)	
18-21	57,308 (55.3%)
22-25	22,554 (21.8%)
26+	21,723 (20.9%)
N3Q72: Year in school? — freshman 1, sophomore 2, junior 3, senior 4, 5th+ year undergrad 5, masters 6, doctorate 7, not seeking a degree 8	
Freshman	20,768 (20.0%)
Sophomore	17,218 (16.6%)
Junior	20,594 (19.9%)
Senior	15,795 (15.2%)
5th+ year undergrad	4,779 (4.6%)
Masters	13,641 (13.2%)
Doctorate	8,197 (7.9%)
Not seeking degree	301 (0.3%)
Other	1,092 (1.1%)
N3Q75A1-8: Race/ethnicity?	
American Indian/Native Alaskan	2,354 (2.3%)
Asian/Asian American	17,411 (16.8%)
Black/African American	6,918 (6.7%)
Hispanic/Latina/Latino	19,113 (18.4%)
Middle Eastern/North African/Arab	1,985 (1.9%)
Native Hawaiian/Pacific Island	615 (0.6%)
White	62,782 (60.6%)
Biracial/multiracial	5,183 (5.0%)
N3Q77A: Are you a member of a fraternity or sorority? — no 1, yes 2	
No	95,633 (92.2%)
Yes	6,656 (6.4%)
N3Q80: Approximate GPA? – A+ to A- (1 to 3) were summarized as A, B+ to B- (4 to 6) were summarized as B, C+ to C- (7 to 9) were summarized as C, D+ to F (10 to 13) were summarized as D/F	
A	63,718 (63.7%)
B	29,899 (29.9%)
C	5,841 (5.8%)
D/F	551 (0.6%)

Table 2: Alcohol-Related Behaviors

Study Sample Size	103,639
N3Q22B2: Frequency of alcohol use in last 3 months — never 0, once or twice 2, monthly 3, weekly 4, daily/almost 6	
Never	6,158 (5.9%)
Once or twice	22,085 (21.3%)
Monthly	18,418 (17.8%)
Weekly	22,551 (21.8%)
Daily/almost daily	1,725 (1.7%)
N3Q22BL2: Have you experienced health, social, legal, and/or financial problems from alcohol? last 3 months — never 0, once or twice 4, monthly 5, weekly 6, daily/almost 7	
Never	57,649 (55.6%)
Once or twice	4,940 (4.8%)
Monthly	1,137 (1.1%)
Weekly	777 (0.7%)
Daily/almost daily	181 (0.2%)
N3Q22BM2: Have you failed to do what's normally expected of you because of alcohol? last 3 months — never 0, once or twice 5, monthly 6, weekly 7, daily/almost 8	
Never	58,603 (56.5%)
Once or twice	4,768 (4.6%)
Monthly	770 (0.7%)
Weekly	470 (0.5%)
Daily/almost daily	111 (0.1%)
N3Q22BN2: Has anyone expressed concern about your alcohol use? — never 0, yes in last 3 months 6, yes but not last 3 months 3	
Never	63,272 (61.0%)
Yes, but not last 3 months	4,456 (4.3%)
Yes, in last 3 months	3,030 (2.9%)
N3Q22BO2: Have you ever tried and failed to control/cut down/stop your alcohol consumption? — never 0, yes in last 3 months 6, yes but not last 3 months 3	
Never	64,796 (62.5%)
Yes, but not last 3 months	2,728 (2.6%)
Yes, in last 3 months	3,272 (3.2%)
N3Q25B1: The last time you drank, did you get drunk? — no 1, yes 2	
No	42,796 (41.3%)
Yes	23,981 (23.1%)
N3Q25B2: The last time you drank, did you INTEND to get drunk? — no 1, yes 2	
No	42,103 (40.6%)
Yes	24,598 (23.7%)
N3Q28: In the last 2 weeks, how many times have you binge drank? — none 1, 1x 2, 2x 3, 3x 4, 4x 5, 5x 6, 6x 7, 7x 8, 8x 9, 9x 10, 10+ 11	
none	24,530 (23.7%)
1x	10,960 (10.6%)
2x	5,494 (5.3%)
3x	2,119 (2.0%)
4x	1,315 (1.3%)

5x	514 (0.5%)
6x	262 (0.3%)
7x	127 (0.1%)
8x	87 (0.1%)
9x	53 (0.1%)
10x or more	172 (0.2%)
N3Q29A: When drinking alcohol in last 12mo, did you do something you later regretted? — no 1 yes 2	
No	60,907 (58.8%)
Yes	13,062 (12.6%)
N3Q29B: When drinking alcohol in last 12mo, did you blackout? (complete memory loss while drunk) — no 1 yes 2	
No	66,541 (64.2%)
Yes	7,477 (7.2%)
N3Q29C: When drinking alcohol in last 12mo, did you brownout? (partial memory loss while drunk) — no 1 yes 2	
No	59,744 (57.6%)
Yes	14,299 (13.8%)
N3Q29D: When drinking alcohol in last 12mo, did you get in trouble with police? — no 1 yes 2	
No	73,364 (70.8%)
Yes	595 (0.6%)
N3Q29E: When drinking alcohol in last 12mo, get in trouble with college authorities? — no 1 yes 2	
No	73,384 (70.8%)
Yes	557 (0.5%)
N3Q29F: When drinking alcohol in last 12mo, did you someone have sex with you without consent? — no 1 yes 2	
No	73,063 (70.5%)
Yes	870 (0.8%)
N3Q29G: When drinking alcohol in last 12mo, did you have sex with someone without consent? — no 1 yes 2	
No	73,843 (71.2%)
Yes	180 (0.2%)
N3Q29H: When drinking alcohol in last 12mo, did you have unprotected sex? — no 1 yes 2	
No	66,468 (64.1%)
Yes	7,535 (7.3%)
N3Q29I: When drinking alcohol in last 12mo, did you physically injure yourself? — no 1 yes 2	
No	69,572 (67.1%)
Yes	4,382 (4.2%)
N3Q29J: When drinking alcohol in last 12mo, did you physically injure another person? — no 1 yes 2	
No	73,591 (71.0%)
Yes	390 (0.4%)
N3Q29K: When drinking alcohol in last 12mo, have you seriously considered suicide? — no 1 yes 2	
No	72,349 (69.8%)
Yes	1,654 (1.6%)

N3Q29L: When drinking alcohol in last 12mo, did you need medical help? — no 1 yes 2	
No	73,168 (70.6%)
Yes	676 (0.7%)
N3Q30A: Last 30 days, did you ever drive after drinking? — no 1, yes 2	
No	39,167 (37.8%)
Yes	5,747 (5.5%)
N3Q30B: Last 12 months, alcohol affecting academics — no 1, negative impact on class performance 2, delayed degree progress 3	
No	72,029 (69.5%)
Negative impact on class performance	1,627 (1.6%)
Delayed degree progress	355 (0.3%)
ALCOHOLRISK: Risk of alcohol misuse, where a score of 0-10 indicates low risk, 11-26 indicates moderate risk, and 27-39 indicates high risk	
Low risk	57,348 (55.3%)
Moderate risk	9,900 (9.6%)
High risk	985 (1.0%)

Table 3: Mental Health and General Health

Study Sample Size	103,639
N3Q1: Overall health rating — Excellent is 1, Very Good 2, Good 3, Fair 4, Poor 5	
Excellent	11,154 (10.8%)
Very Good	36,483 (35.2%)
Good	35,706 (34.4%)
Fair	10,968 (10.6%)
Poor	1,497 (1.4%)
N3Q48: Stress level, last 30 days — no stress 1, low 2, moderate 3, high 4	
No stress	1,861 (1.8%)
Low	22,322 (21.5%)
Moderate	51,360 (49.5%)
High	26,931 (26.0%)
N3Q65A2: Professional diagnosis on ADD/ADHD? — no 1, yes 2	
No	87,472 (84.3%)
Yes	14,032 (13.5%)
N3Q65A3: Professional diagnosis on alcohol or drug abuse/addiction? — no 1, yes 2	
No	100,041 (96.4%)
Yes	1,568 (1.5%)
N3Q65A7: Professional diagnosis on anxiety? — no 1, yes 2	
No	65,895 (63.6%)
Yes	35,840 (34.6%)
N3Q65A15: Professional diagnosis on depression? — no 1, yes 2	
No	74,446 (71.8%)
Yes	27,210 (26.2%)
N3Q65A19: Professional diagnosis on gambling disorder? — no 1, yes 2	
No	101,144 (97.6%)
Yes	162 (0.2%)
N3Q65A28: Professional diagnosis on insomnia? — no 1, yes 2	
No	93,560 (90.2%)
Yes	7,239 (7.0%)
N3Q65A31: Professional diagnosis on OCD? — no 1, yes 2	
No	95,464 (92.1%)
Yes	6,133 (5.9%)
N3Q65A33: Professional diagnosis on PTSD? — no 1, yes 2	
No	93,279 (90.0%)
Yes	8,292 (8.0%)
N3Q65A35: Professional diagnosis on sleep apnea? — no 1, yes 2	
No	98,852 (95.3%)
Yes	2,394 (2.3%)
N3Q65Y: How much have your selected conditions negatively affected academics? (30 days) — did not affect 1, negatively impacted class progress 2, delayed degree progress 3	
Did not affect	46,801 (45.1%)
Negatively impacted class progress	23,399 (22.6%)

Delayed degree progress	5,332 (5.1%)
RKESSLER6: Kessler 6 Screening for Non-Specific Serious Mental Illness Score Collapsed, 0-12 negative for serious psychological distress is 1, 13-24 positive is 3	
Negative for serious psychological distress	81,341 (78.5%)
Positive for serious psychological distress	19,708 (19.0%)
RULS3: UCLA Loneliness Scale Score Collapsed, 3-5 negative for loneliness is 1, 6-9 positive is 2	
Negative for loneliness	52,660 (50.8%)
Positive for loneliness	49,609 (47.9%)
RSBQR: Suicide Behavior Questionnaire-Revised (SBQR) Screening Score, 3-6 negative suicidal screening, 7-18 positive suicidal screening	
Negative suicidal screening	75,542 (72.8%)
Positive suicidal screening	26,377 (25.4%)
DIENER: Diener Flourishing Scale – Psychological Well-Being (PWB) Score (8-56), with higher scores reflecting higher PWB.	
8-16	1,471 (1.4%)
17-24	1,738 (1.7%)
25-32	3,134 (3.0%)
33-40	11,813 (11.4%)
41-48	22,978 (22.2%)
49-56	30,160 (29.1%)
CDRISC2: Connor-Davison Resilience Scale (CD-RISC) (0-8), with higher scores reflecting greater resilience	
0-2	2,356 (2.3%)
3-5	29,067 (28.0%)
6-8	69,248 (66.8%)

3.2 Data Cleaning and Preprocessing

Another important step in utilizing a data science pipeline, beyond just collecting the data and identifying variables, response types, and other elements, is cleaning and preprocessing the data to be fit for analysis. In this case, there were a couple rounds of cleaning. There was an initial cleaning, which involved removing variables that seemed largely irrelevant to this study, and then another wave of variable removal to further narrow the focus. Many of the survey questions used in this study included preset answers for the students to respond to, such as Yes/No questions or Likert scale questions, where the responses range from no stress to high stress, or from very unlikely to very likely, for example. For the purpose of analysis, all responses in the data were converted from their original categorical form, such as very likely, to their numerical values, such as 5, which exist in the code-book of the data. This process was significantly expedited by a JSON file that contained the numeric code for each answer to each question in the survey.

When cleaning data, there are sometimes unexpected issues that arise. In converting the data to all numeric responses, there was an issue with question N3Q28, where the students' responses of "None" to the question of binge drinking frequency were being counted as NaN (not a number) values instead. This needed to be fixed, as binge drinking frequency seemed to be a variable that would be very important to this study. Due to all

of these variables being bound to a certain range, no winsorization or removal of outliers was necessary. Some variables are formed from groups of other, more specific variables, with only the summarized versions being used for the sake of this study. These variables are shown below:

- RBMI, which is derived from a person's weight and height, placing them into a category estimating how healthy they are accordingly.

- RKESSLER6, which is the Kessler 6 Screening for Non-Specific Serious Mental Illness, including six questions on mental health struggles and associated distress (found under N3Q44 in the survey)

- RULS3, which is the UCLA Loneliness Scale, including three questions on loneliness (found under N3Q45 in the survey)

- RSBQR, which is the Suicide Behavior Questionnaire-Revised, including based on several questions in the survey regarding suicide risk (found at N3Q49 - N3Q52 in the survey)

- DIENER, which is the Diener Flourishing Scale, representing a group of questions on overall psychological well-being (found under N3Q41 in the survey)

- CDRISC2, which is the Connor-Davison Resilience Scale, indicating the extent of someone's resilience from a group of survey questions (found under N3Q42 in the survey)

- ALCOHOLRISK, which utilizes questions from the Alcohol, Smoking and Substance Involvement Screening Test (found at N3Q22A - N3Q22Q in the survey) to address the level of vulnerability someone has to hazardous drinking, harmful use, alcohol dependence, and other consequences

For clear reporting purposes, profiling results were presented before cleaning results in the body of this text, though in practice, cleaning was performed first to allow demographic statistics and other tables and visualizations to be untarnished by duplicates and other issues. As previously stated, missing values were addressed on a case-by-case basis, rather than eliminating entire rows just because of one missing value, for example. Refer to Appendix A: Cleaning for the Python code used for this section of the pipeline.

3.3 Data Analysis and Visualization

As the study moved into the exploratory data analysis (EDA) and visualization phase, some variables were not utilized in order to narrow the focus of the study and prepare for the modeling process that was to follow. For example, in the tables in the Data Collection and Profiling section, there are 24 alcohol-related variables alone, which is an overwhelming amount for analysis, visualization, and modeling. Some of these variables may be reintroduced in the modeling section, but the analysis and visualization will focus primarily on how alcohol relates to health-related factors and the most general alcohol-related variables, such as the ALCOHOLRISK variable, which effectively summarizes the majority of the 24 alcohol-related variables in the aforementioned table. Table 4 shows

the variables that will be used in this section. Note that some of these variables were added in later stages of analysis.

Table 4: Variables Used in EDA and Visualization

Variable	Description
Alcohol-Related Factors	
N3Q22B2:	Frequency of alcohol use in the last 3 months
N3Q28:	Frequency of binge drinking in the last 2 weeks
ALCOHOLRISK:	Risk of alcohol misuse, 0-10 indicates low risk, 11-26 indicates moderate risk, and 27-39 indicates high risk
Demographics	
N3Q67A:	Sex
N3Q72:	Year in college
N3Q77A:	Is the student a member of a fraternity or sorority?
N3Q80:	Student GPA
Health-Related Factors	
N3Q1:	Self-reported overall health rating
N3Q48:	Stress level in the last 30 days
N3Q65A2:	Has the student been diagnosed with ADD/ADHD?
N3Q65A7:	Has the student been diagnosed with anxiety?
N3Q65A15:	Has the student been diagnosed with depression?
N3Q65A19:	Has the student been diagnosed with a gambling disorder?
N3Q65A28:	Has the student been diagnosed with insomnia?
N3Q65A31:	Has the student been diagnosed with OCD?
N3Q65A33:	Has the student been diagnosed with PTSD?
N3Q65A35:	Has the student been diagnosed with sleep apnea?
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)
RSBQR:	Suicide Behavior Questionnaire-Revised score, 3-6 indicates negative suicidal screening, 7-18 indicates negative suicidal screening
DIENER:	Diener Flourishing Scale (8–56; higher = greater well-being)
CDRISC2:	Connor-Davison Resilience Score (0-8), with higher scores reflecting greater resilience

Principal component analysis (PCA) is a statistical technique used to reduce redundancy and noise among similar variables by combining them into a set of new, artificial variables. Instead of applying PCA at this point in the pipeline, variables were manually selected for the EDA and analysis phase. This approach was chosen to reduce overlap between certain survey responses while still maintaining nuance in certain areas, such as being able to see how specific mental disorders relate to alcohol use. The choice and justification for not utilizing PCA will be addressed further in the model building and testing section.

In terms of how demographics relate to alcohol use and misuse (drinking in a way that is harmful to oneself or others), there were some interesting findings. Between male and female students, there were few notable factors. Males did drink a little more overall, where females had drank alcohol once, twice, or monthly more often than males, but males had drank alcohol weekly or almost daily more often than females. This time window was the last 3 months before the student took the survey. Males and females binge drank around the same amount in the 2 weeks before taking the survey; however, males had a slightly higher proportion in the range of moderate to high risk drinkers, at around 18% of males compared to 15.1% of females. Moving on to year in school, the proportion of students who were in the moderate to high range for risk of alcohol misuse remained quite consistent, with the smallest groups being doctorate students and freshman students at 13.4% and 14.7%, respectively, whereas the largest

groups were represented by students who stated they were not seeking a degree at 18.7%, 5th year undergraduates at roughly 17.6%, and senior students at 17.2%. Binge drinking rates and overall alcohol use rates remained steady across these groups of students as well.

The more interesting findings from this demographics section, however, are those related to GPA and fraternity/sorority participation. For GPA, the proportion of students who were at moderate to high risk of alcohol misuse increased as GPA decreased, indicating that students with higher GPAs tend to more often be part of the group at low risk. Figure 2 displays a stacked bar chart showing this relationship.

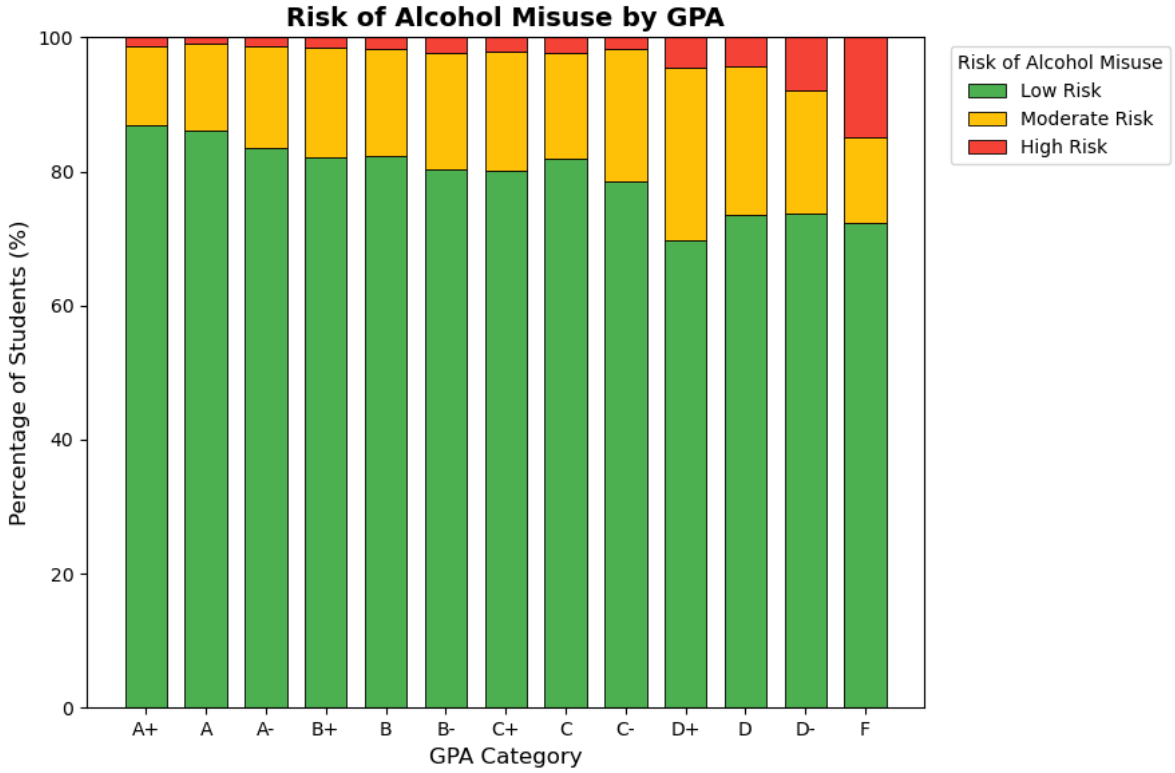

Figure 2: The risk of alcohol misuse by GPA.

There were no impactful correlations between GPA and binge drinking or alcohol consumption rates, suggesting that frequency of drinking is less important to a student's academic success than the risk factors associated with drinking, as seen by the correlation between higher chances of misuse among students with lower GPAs.

For students that said they were part of a fraternity or sorority ("Greek life"), the rates of binge drinking stayed roughly the same yet again. However, alcohol was consumed more often by students in Greek life and these students also had a notably higher risk of alcohol misuse. For students involved in Greek Life, 73.7% were at low risk, 24.2% were at moderate risk, and 2.15% were at high risk, compared to 84.9% for low risk, 13.7% at moderate risk, and 1.4% at high risk for those not involved in Greek Life. An additional variable was added here for this section that was not initially mentioned, that being question N3Q77B which asks the student if they *live* in a fraternity or sorority, not just participate in one. For these students that live in a fraternity or sorority, alcohol consumption rates are even higher, which on its own isn't necessarily a problem, however, the proportions of risk of alcohol misuse are also higher. 65.1% of students who live in Greek life consume alcohol either weekly or daily/almost daily, compared

to 33.4% of students who do not live in Greek housing. Figure 3 displays a stacked bar chart showing the relationship between general alcohol consumption and whether a student lives in a fraternity or sorority.

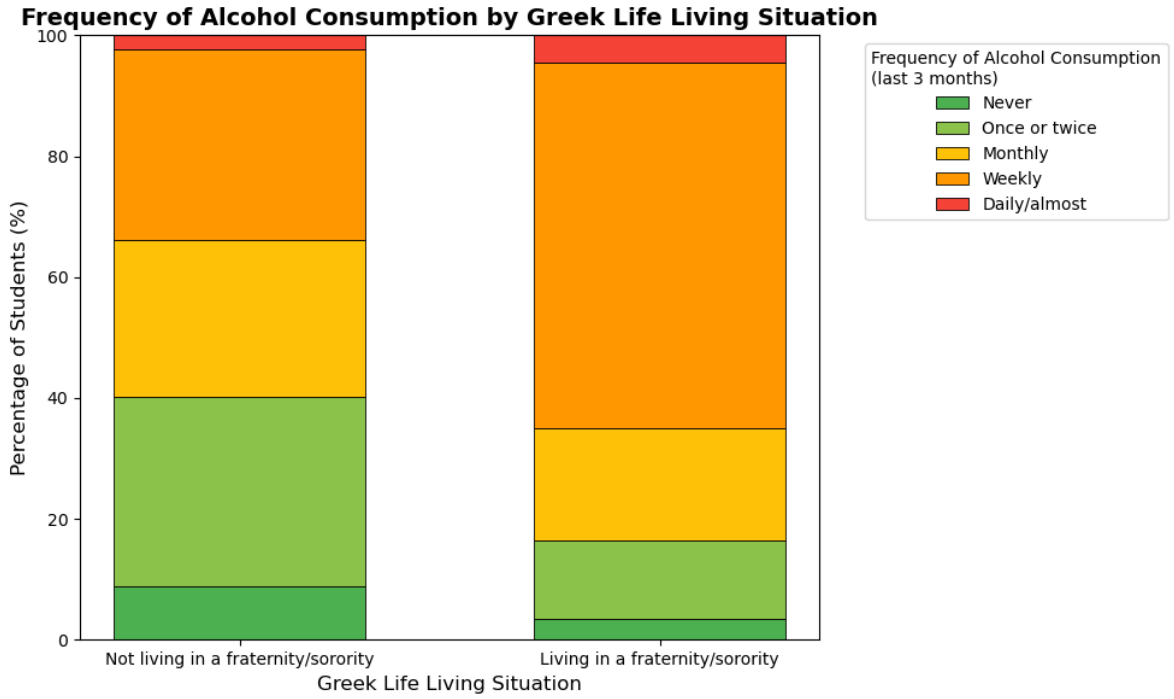

Figure 3: Frequency of alcohol consumption by Greek life living situation.

Beyond general consumption rates however, only 15.7% of students who do not live in Greek housing fall in the moderate to high risk range of alcohol misuse, whereas 32.6% of students who do live in Greek housing fall in this range. This highlights the importance of educating fraternity and sorority members on responsible alcohol consumption and associated behaviors, as these higher numbers apply to both those who are simply a member of the organization, and also those who live in the house itself. Figure 4 displays a stacked bar chart showing the relationship between these categories.

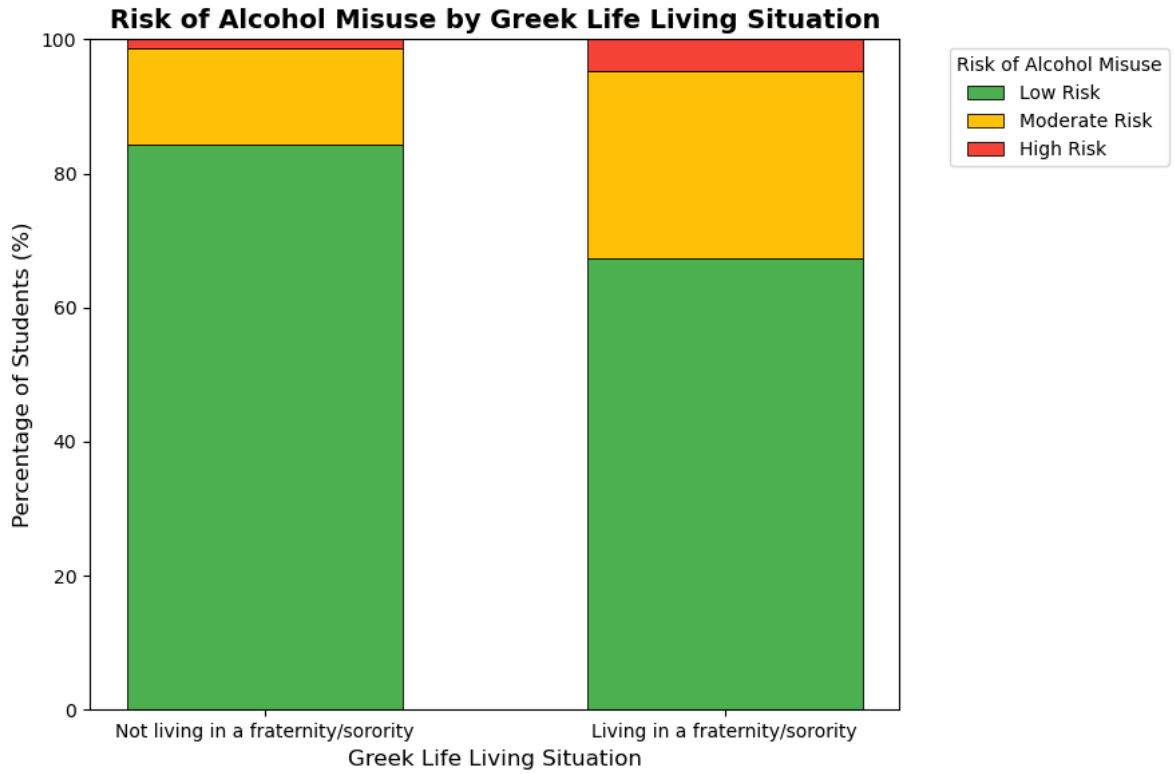

Figure 4: The risk of alcohol misuse by Greek life living situation.

In terms of seeing how alcohol relates to health-related factors, there were some notable findings as well. The mental and behavioral disorders being investigated in this study are ADD/ADHD, anxiety, depression, gambling disorder, insomnia, OCD, PTSD, and sleep apnea. If a student has at least one of any of these disorders, they are more likely to have a moderate-high risk of alcohol misuse than those without the same disorder. Table 5 summarizes these findings.

Table 5: Mental and Behavioral Disorders and Alcohol Misuse Risk

Disorder	**Moderate–High Alcohol Risk (%) Disorder Absent**	**Moderate–High Alcohol Risk (%) Disorder Present**	**Difference**
ADD/ADHD	15.09	20.33	+5.24%
Anxiety	14.42	18.20	+3.78%
Depression	14.18	19.73	+5.55%
Gambling Disorder	15.83	51.09	+35.26%
Insomnia	15.54	19.94	+4.40%
OCD	15.58	20.29	+4.71%
PTSD	15.40	20.61	+5.21%
Sleep Apnea	15.80	18.23	+2.43%

Note. Moderate–High alcohol misuse risk includes students classified as moderate or high risk according to the ALCOHOLRISK composite. Disorder presence is based on self-reported clinical diagnosis (N3Q65 question series).

From Table 5, it is apparent that the presence of any of these common mental or behavioral disorders suggest around a 4.5-5% increased chance that a student is a moderate to high risk drinker. This is not causal, of course, as this is possibly a result of students with these disorders potentially drinking more often or having harmful drinking behaviors, which actually leads to

the increased risk of alcohol misuse. Note that the abnormally large chance of alcohol misuse for students with a gambling disorder may be due to the sample size of that population, with it being only 162 students, while these other populations range from 2,394 students to 35,840.

Further exploring the relationships between disorders and alcohol-related factors, there were no significant findings in the presence of disorders correlating with binge drinking rates, nor were there between the presence of disorders and general alcohol use rates. For stress, there were mild relationships. For example, students with high stress had the biggest proportion in the moderate to high risk group with 19.5% of high stress students being in this range, compared to only 12% of low stress students, the smallest proportion in the moderate to high risk range. Binge drinking, yet again, showed no real trend with the other variable at play, this time being stress. There was little to no correlation between stress level and general alcohol consumption as well. Much of the same can be said when comparing alcohol-related factors to the self-reported overall health ratings of students, where the most notable correlation was involving risk of alcohol misuse. 22.2% of students who report having poor overall health fell in the range of moderate to high alcohol risk, whereas, on the opposite end only 12.8% of students who report excellent health fall into this moderate to high risk range. This being said, however, the proportions of students who either didn't drink or only drank once or twice during the 3 month period actually decreased as overall health increased towards excellent (16.2% to 8.2% and 34.7% to 28.7%, respectively), whereas the proportions of students who drank weekly to monthly increased as overall health increased (22.9% to 34.7% and 21.4% to 25.7%, respectively). This suggests that consuming alcohol, but consuming it responsibly in weekly to monthly moderation is not harmful and may even be beneficial for some students.

Moving into the collapsed mental health scores, students who scored positive in mental distress (RKESSLER6), loneliness (RULS3), and/or suicidal behavior (RSBQR) were also more likely to be at moderate or high risk for alcohol misuse. Lower DIENER scores, which indicates lower quality of well-being also showed higher numbers in those moderate-to-high risk range than those with a higher quality of life. For students with lower resilience (CDRISC2), the same was also true. Table 6 summarizes the results of these collapsed mental health scores.

Table 6: Mental Health Scores and Alcohol Misuse Risk

Scale	**Moderate–High Alcohol Risk Not having mental health issue**	**Moderate–High Alcohol Risk Having mental health issue**	**Difference**
RKESSLER6 (mental distress)	13.98%	23.82%	+9.84%
RULS3 (loneliness)	13.27%	18.84%	+5.57%
RSBQR (suicidal behavior)	13.29%	22.42%	+9.13%

Comparing the results from table 5 to table 6, it currently seems that the presence of issues like mental distress and suicidal behavior are slightly more impactful on moderate to high alcohol misuse risk than the diagnosed mental disorders themselves (anxiety, depression, etc.) Furthermore, however, are the DIENER and CDRISC2 scores. As DIENER scores increase (as quality of life increases), interestingly there are less students drinking no alcohol, but also less students consuming alcohol daily/almost daily, again that alcohol consumed in moderation is beneficial for some students' well being, showed similarly in the previous discussion of overall health rating. With binge drinking, yet again, there were no notable correlations. Alcohol misuse risk showed similar findings to the majority of the other relationships investigated here, were less desirable outcomes were correlated with moderate to high risk of alcohol misuse. Figure 5 showcases this relationship between the DIENER scale and alcohol risk, where around 30% at the lower DIENER scores are at moderate to high risk for alcohol related issues compared to only around 15% at the higher DIENER scores.

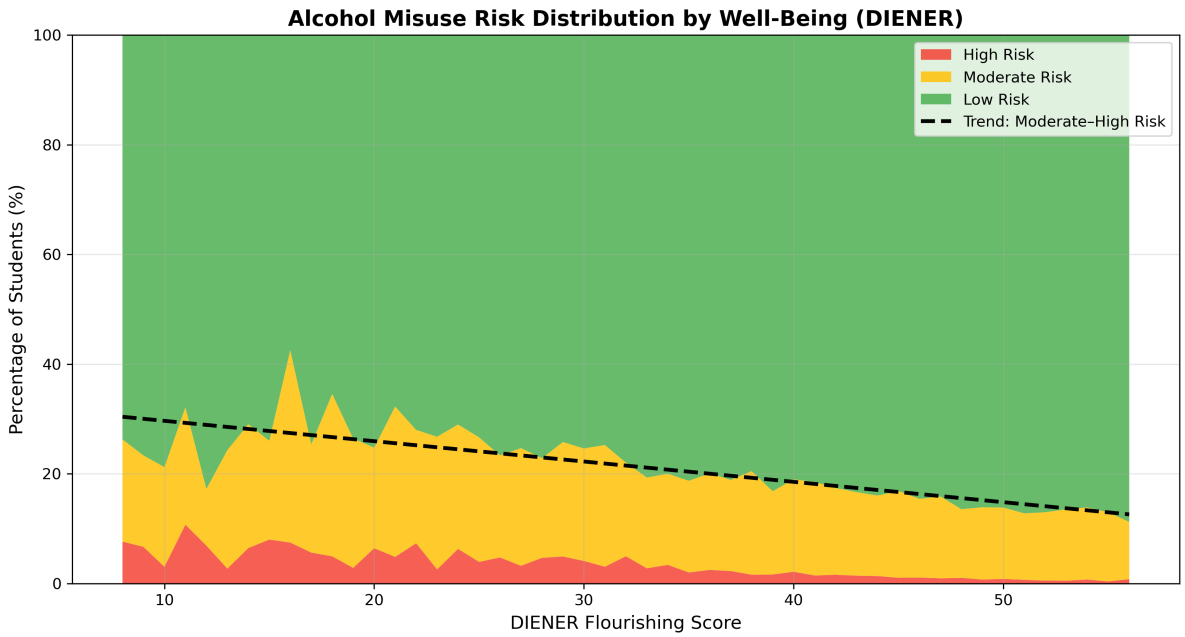

Figure 5: The risk of alcohol misuse by DIENER score (higher scores equal greater well-being).

Much of the same that was said for DIENER scores can be said for the CDRISC2 scores as well, which measure a student's resilience on a scale of 0-8, where 8 indicates the highest and most ideal level of resilience. Students with lower resilience scores tended to have a greater chance of being in the moderate to high risk for misuse zone, with 31.9% at the lowest resilience rating down to 13.7% at the highest resilience rating.

To explore additional relationships, some more alcohol-related variables were incorporated. These were chosen as they offered slightly more specific insights into alcohol-related issues. These include:

- N3Q22BL2: Have you experienced health, social, legal, and/or financial problems from alcohol in the last 3 months?

- N3Q22BO2: Have you ever tried and failed to control/cut down/stop your alcohol consumption?

- N3Q29B: When drinking alcohol in the last 12 months, did you blackout?

- N3Q30A: In the last 3 months, did you ever drive after drinking?

Table 7 summarizes the findings of these newly implemented questions as they relate to GPA.

Table 7: Alcohol-Related Issues and GPA

Question	**% of Students with Outcome**				**Avg. Rate of Change**
	A range	**B range**	**C range**	**D/F range**	
N3Q22BL2: Health, social, legal, or financial problems from alcohol	9.7%	13.2%	13.1%	20.2%	+3.5%
N3Q22BO2: Failed attempts to control, cut down, or stop alcohol use	7.9%	9.5%	10.4%	17.3%	+3.1%
N3Q29B: Experienced blackout while drinking	8.9%	12.2%	13.6%	20.6%	+3.9%
N3Q30A: Drove after drinking alcohol	13.2%	12.4%	14.7%	15.1%	+0.63%

As seen by the GPA table, the size of the group with negative outcomes increases as GPA decreases, typically by around 3.5% with each subsequent GPA decrease, indicating that students with lower GPAs are typically those at greater risk of alcohol related harm, and vice versa. This increase, however, is mostly non-existent for the question of driving after the influence, meaning that the rates of students who drive under the influence is relatively consistent across GPA.

Moving back to the conversation on Greek housing situation, see Table 8 and Figure 6 below which both summarize the relationships between these questions and the Greek living situation.

Table 8: Alcohol-Related Issues and Greek Life Living Situation

Question	**% of Students with Outcome**		**Difference**
	Not in Greek housing	**In Greek housing**	
N3Q22BL2: Health, social, legal, or financial problems from alcohol	10.7%	23.8%	+13.1%
N3Q22BO2: Failed attempts to control, cut down, or stop alcohol use	8.4%	10.6%	+2.2%
N3Q29B: Experienced blackout while drinking	9.9%	26.7%	+16.8%
N3Q30A: Drove after drinking alcohol	12.8%	11.6%	-1.2%

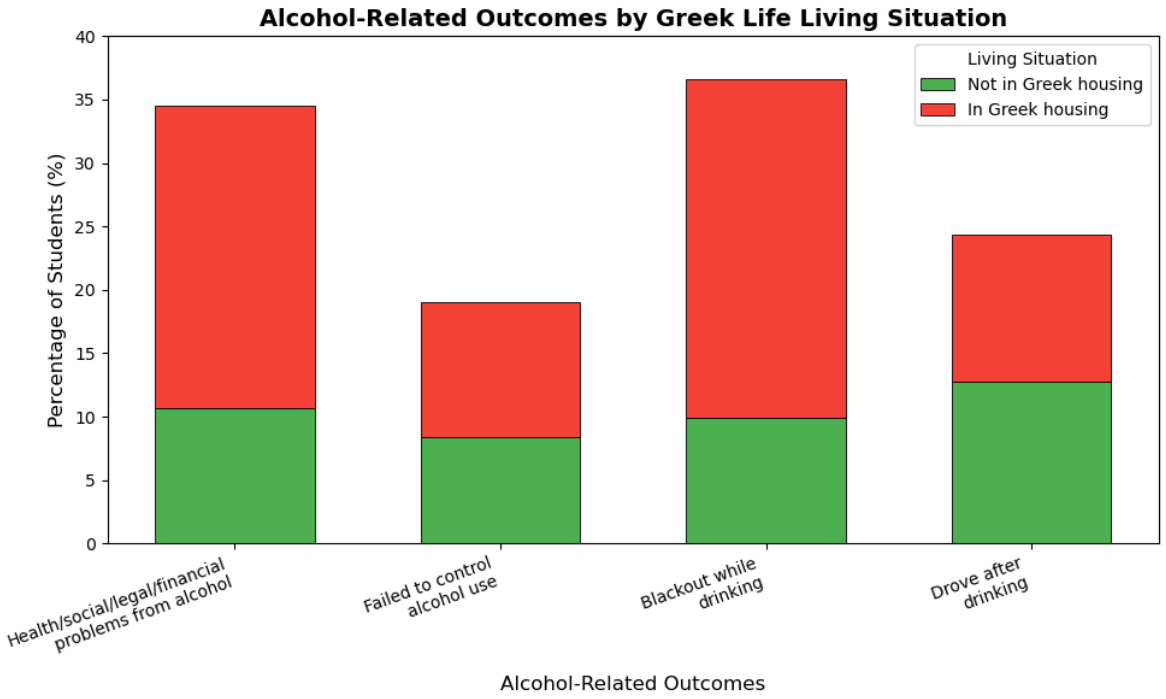

Figure 6: Alcohol-Related Outcomes by Greek Life Living Situation).

This table offers some very interesting insights. Students who live in Greek housing are significantly more likely to experience health, social, legal, and financial problems from alcohol, and are also significantly more likely to blackout when drinking, which comes with its own issues such as hindered cognitive abilities like worsened impulse control, judgment, and decision-making which leads to a greater chance of injuries and sexual assault. However, in contrast, students in Greek housing actually drink and drive less than the rest of the college population. This is only a 1.2% difference, but is still an interesting finding here. Overall, students who live in Greek housing have a greater risk of alcohol-related harm than those who simply participate in Greek life, and the group who just participates already tends to be a little higher risk than those who are not involved in Greek life at all, as addressed previously. All this to say that education on risks of blackout drinking and alcohol-related consequences is very important for those involved in Greek Life on a college campus.

Moving on to the collapsed mental health-score variables such as RKESSLER6, RULS3, and RSBQR, Table 9 below displays the relationships between these metrics and the questions currently being discussed.

Table 9: Alcohol-Related Issues and Mental Health Issues

Question	**% of Students with Outcome**		
	SPD (RKESSLER6)	**Loneliness (RULS3)**	**Suicidal behavior (RSBQR)**
N3Q22BL2: Health, social, legal, or financial problems from alcohol	16.7%	13.4%	14.7%
N3Q22BO2: Failed attempts to control, cut down, or stop alcohol use	13.4%	10.2%	13.4%
N3Q29B: Experienced blackout while drinking	13.5%	11.3%	13.3%
N3Q30A: Drove after drinking alcohol	13.1%	13.8%	16.2%

Regarding this table, a larger proportion of students with any negative drinking outcome also had a negative mental health outcome than students with the same drinking outcome but without the corresponding negative mental health outcome, whether that be serious psychological distress (SPD), loneliness, or suicidal behavior. In all 12 cases displayed, having SPD, loneliness, or suicidal behavior placed students at greater risk of negative alcohol-related outcomes, with differences in proportions remaining relatively consistent: the smallest difference was 0.4% more students driving after drinking with SPD than without, and the largest was 7.3% more students experiencing health, social, legal, or financial problems due to alcohol with SPD than those without. Overall, loneliness had the weakest effect, followed by suicidal behavior, which was then barely preceded by SPD, though each of these variables exhibited subtle but consistent and prevalent effects on risk of alcohol-related harm. As expected, students positive for the mental health issues in the collapsed scores, as well as those with lower scores for DIENER and CDRISC2, also experienced more alcohol-related problems. RKESSLER6 showed the largest difference in collapsed-score groups, with 9.4% of students without SPD experiencing these issues compared to 16.7% of students with SPD, a 7.3% difference. For both DIENER well-being scores and CDRISC2 resilience scores, the proportion of students who experienced alcohol-related problems ranged roughly from 20% for low scores to 8% for high scores, indicating that students who tend to experience fewer alcohol-related problems also have a greater quality of life and higher resilience.

Finally, Tables 10 and 11 display the relationships between certain mental and behavioral disorders and negative alcohol-related outcomes.

Table 10: Alcohol-Related Issues and Mental/Behavioral Disorders (Part 1)

Question	ADHD	Anxiety	Depression	Gambling Disorder
N3Q22BL2: Health, social, legal, or financial problems from alcohol	13.2%	11.8%	12.6%	48.2%
N3Q22BO2: Failed attempts to control, cut down, or stop alcohol use	13.2%	11.1%	12.7%	42.1%
N3Q29B: Experienced blackout while drinking	11.6%	11.0%	11.7%	33.3%
N3Q30A: Drove after drinking alcohol	14.8%	13.0%	14.4%	27.1%

Table 11: Alcohol-Related Issues and Mental/Behavioral Disorders (Part 2)

Question	Insomnia	OCD	PTSD	Apnea
N3Q22BL2: Health, social, legal, or financial problems from alcohol	14.0%	14.0%	13.7%	11.8%
N3Q22BO2: Failed attempts to control, cut down, or stop alcohol use	14.9%	13.8%	16.8%	14.4%
N3Q29B: Experienced blackout while drinking	11.4%	13.0%	11.9%	10.0%
N3Q30A: Drove after drinking alcohol	14.3%	13.8%	14.4%	19.7%

From these two tables, it is apparent that gambling disorder had by far the strongest indication of contributing to negative alcohol-related outcomes. As stated previously, the sample size of students with a gambling disorder is only 162, compared to 2,300 to 35,000 for all of the other disorders here. This small sample size may be contributing to the inflated numbers for gambling disorder here. However, gambling disorder does overlap with impulsivity, poor self-regulation, and possibly more frequent substance use, so there may be some truth here. However, in addition to gambling disorder, all of the outcomes of other disorders sit between the proportions of 10% to 19.7%, meaning that, for example, 10% of students with sleep apnea experienced blackout while drinking. Note that the proportions of students with each disorder who did not experience these problems were not included in this table, as in every single case, having any of these disorders increased the chances of experiencing negative alcohol-related issues. Not considering gambling disorder, the smallest difference in proportions was an increase of 0.3% between students with anxiety who did not drive under the influence and those who did, and the largest difference was an increase of 9.2% between students with PTSD who tried and failed at reducing alcohol intake, and those who did not try and fail at reducing alcohol intake. Looking at health from a larger scale, the question of a student's self-reported overall health is not included in a table here, but the size of this group who had experienced alcohol related problems grew consistently from 8.4% to 17% as students reported worse overall health ratings as well.

The next section of this report will discuss the creation and testing of a predictive model, which will incorporate the insights of this analysis and visualization section to attempt to offer the highest quality predictive power.

3.4 Model Building and Testing

After reviewing assorted literature on the topic and performing analysis firsthand, a decision was made to develop an ordinal logistic regression model. A standard logistic regression model focuses on two outcomes, whereas ordinal logistic regression models handle outcomes with a natural order. The variable "ALCOHOLRISK" will be the focus of the model, where the ordinal logistic regression will attempt to classify a student as having a low, moderate, or high risk of alcohol misuse based on factors such as the diagnosis of mental health conditions, the collapsed mental health scores (RKESSLER6, RSBQR, etc), and other metrics. Logistic regression and other forms of regression modeling were used in many of the studies covered in the literature review.

Before testing the model, however, some preliminary work had to be done. Missing values were cleared from every column that would used in the modeling process. This resulted in a total of 58,933 rows as the sample size for the modeling process. Many columns were reformatted for the sake of consistency and modeling standards. Examples of this include changing all of the numeric coding for all the yes/no questions from no being 1 and yes being 2 to no being 0 and yes being 1, as well as scaling the DIENER and CDRISC2 variables, which sets their mean equal to 0 and standard deviation equal to 1 to aim for more balanced model performance. Next, variables were selected for the model. Table 12 below displays these variables.

Table 12: Variables Used in Ordinal Logistic Regression

Variable	**Description**
Target	
ALCOHOLRISK:	Classify a student's risk of alcohol misuse based on the features below (low risk, moderate risk, or high risk)
Mental Health Diagnoses	
N3Q65A2:	Has the student been diagnosed with ADD/ADHD?
N3Q65A7:	Has the student been diagnosed with anxiety?
N3Q65A15:	Has the student been diagnosed with depression?
N3Q65A28:	Has the student been diagnosed with insomnia?
N3Q65A31:	Has the student been diagnosed with OCD?
N3Q65A33:	Has the student been diagnosed with PTSD?
N3Q65A35:	Has the student been diagnosed with sleep apnea?
Mental Health Scales	
RKESSLER6:	Kessler-6 score for mental distress
RULS3:	UCLA Loneliness Scale
RSBQR:	Suicide Behavior Questionnaire-Revised score
DIENER:	Diener Flourishing Scale
CDRISC2:	Connor-Davison Resilience Score
Additional Predictors	
N3Q1:	Self-reported overall health rating
N3Q48:	Stress level in the last 30 days
N3Q77A:	Is the student a member of a fraternity or sorority?
N3Q77B:	Does the student live in a fraternity or sorority?
N3Q80:	Student GPA

Note that the absence of gambling disorder in this model was due to the small sample size in the survey (n = 162) in comparison to the other disorders (between n = 2,394 and n = 35,840) and the correlations between gambling disorder and alcohol consumption that were multiple times higher than the semi-consistent numbers of every other disorder of interest, suggesting higher variance and instability in the data.

After predictors were chosen, collinearity checks were performed to ensure no variables had a prominent enough overlap to bias the model results. Two methods were utilized and were performed on both the mental health diagnosis variables as well as the mental health scale variables. The first of these two methods was a correlation matrix to display general correlation between results from each column of data. Both matrices indicated that there was no overlap significant enough to disrupt model results. Figure 7 below displays this matrix for the mental health scales.

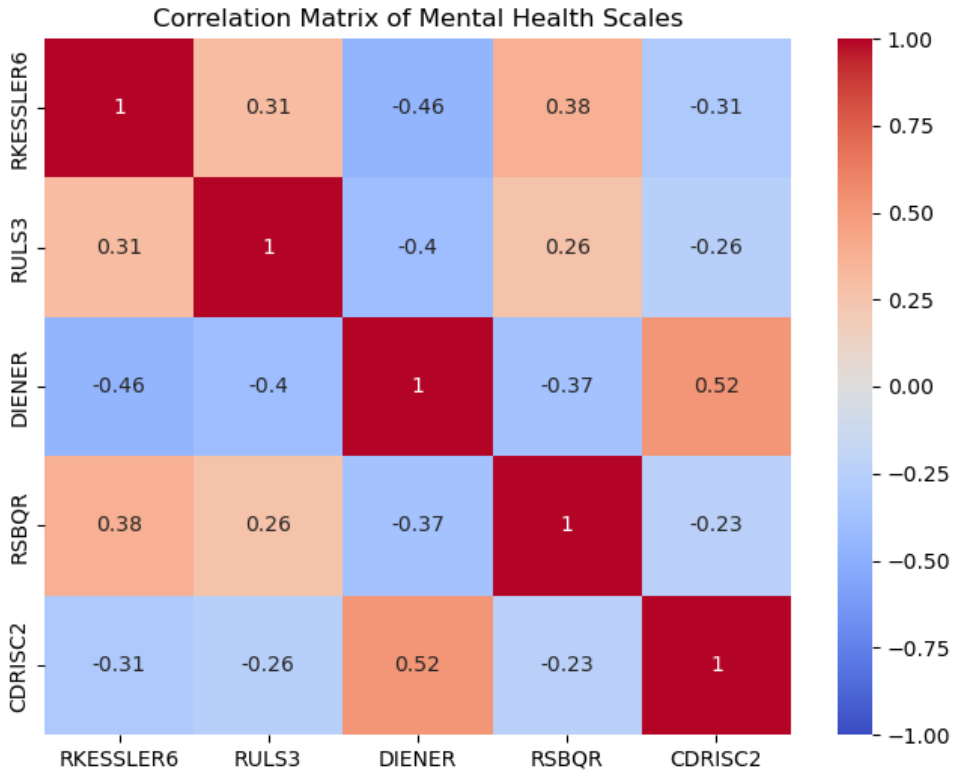

Figure 7: Correlation Matrix of Mental Health Scales .

With the figure above, the goal was to avoid having any values over 0.7 or below -0.7, as above this value would often be considered high collinearity, which is undesirable, as values further from 0 and closer to 1 or -1 make isolating the individual effects of predictors more difficult. As seen in this table, apart from the diagonal of values of 1 which are always generated based on the anatomy of a correlation matrix, there were no values that breached the established threshold. When the same test was run for the diagnoses, results were quite stable as well, although the correlation between anxiety and depression diagnoses was 0.65. In addition to this method, the Variance Influence Factor (VIF) of each of these variables was calculated as well, where a score above 5 would suggest high collinearity. In calculating the VIF, every variable scored under the threshold, with the largest values being 2.65 for anxiety and 2.63 for depression. After checking for collinearity issues, the next step is to establish the parameters of the model itself.

In testing a model, there needs to be split of data for training the model and then additional data for testing the model. This is done to simulate how the model would perform on new, unseen data and allows the model to learn the general patterns of the data. Often times, the

split between training and test data is a 70/30, 75/25, or 80/20 split. This model will utilize a 75/25 split, meaning that 75% of the data (44,199 rows in this case) will be used to train the model by teaching it patterns, and then the remaining 25% (14,734 rows in this case) will be used to test the effectiveness of the model by evaluating its accuracy on new, unseen data.

Upon running the model, some outputs were produced that give insights into the model performance. The first of these is mean absolute error (MAE), which is the average size of the mistakes made by the model's predictions. The MAE score was 0.278, which indicates that, on average, the model's predictions were less than 0.3 categories off, where the jump from low risk to moderate risk is one category, as is the distance from moderate risk to high risk. Additionally, 98.7% of predictions were exact or only off by one category, meaning that only 1.3% of the test data reflected the most severe misclassification (where low risk was predicted to be high risk and vice versa). The observed class imbalance was preserved to reflect the natural distribution of alcohol misuse risk in the population.

These metrics indicate strong performance in maintaining the correct order of risk, but upon further class-specific investigation, it was revealed that there was an imbalance in class sizes between low, moderate, and high alcohol risk students. Due to this, the model predominantly grouped students into the low risk category and sensitivity for moderate and high risk classifications was limited. The best way to interpret this model is that it performs very well at capturing ordinal relationships between predictors and alcohol misuse risk rather than precisely identifying students in higher risk categories.

Furthermore, another output coded into the model was a coefficient table that shows exactly how much each variable contributed to predicting the risk of alcohol misuse. Table 13 shows these values and a brief interpretation of them.

Table 13: Ordinal Logistic Regression Coefficients and Odds Ratios for Predicting Higher Alcohol Risk

Variable (Survey Item)	Coefficient	Odds Ratio	Brief Interpretation
RSBQR – (If a student is positive for suicidal behavior)	0.185	1.204	+20% higher odds of greater alcohol risk
N3Q77A – (If a student is a Greek organization member)	0.181	1.199	+20% higher odds of greater alcohol risk
RKESSLER6 – (If a student is positive for serious for psychological distress)	0.091	1.096	+10% higher odds of greater alcohol risk
RULS3 – (If a student is positive for Loneliness)	0.084	1.087	+9% higher odds of greater alcohol risk
N3Q80 – (GPA, higher values)	-0.085	0.919	-8% lower odds of greater alcohol risk
DIENER – (Well-being scale; standardized; one SD = 8.6 points on the original 8–56 scale)	-0.079	0.924	-8% lower odds of greater alcohol risk
N3Q65A15 – (If a student has been diagnosed with depression)	0.066	1.068	+7% higher odds of greater alcohol risk
N3Q77B – (If a student lives in Greek housing)	0.050	1.051	+5% higher odds of greater alcohol risk
N3Q65A2 – (If a student has been diagnosed with ADHD)	0.044	1.045	+4% higher odds of greater alcohol risk
N3Q48 – (Increase in perceived stress level)	0.043	1.044	+4% higher odds of greater alcohol risk
N3Q65A7 – (If a student has been diagnosed with anxiety)	-0.030	0.970	-3% lower odds of greater alcohol risk
N3Q1 – (Increase in overall health rating)	0.023	1.024	+2% higher odds of greater alcohol risk
N3Q65A33 – (If a student has been diagnosed with PTSD)	0.012	1.012	+1% higher odds of greater alcohol risk
N3Q65A28 – (If a student has been diagnosed with insomnia)	-0.006	0.994	≈ no meaningful association
N3Q65A35 – (If a student has been diagnosed with sleep apnea)	-0.003	0.997	≈ no meaningful association
N3Q65A31 – (If a student has been diagnosed with OCD)	-0.003	0.997	≈ no meaningful association
CDRISC2 – (Resilience scale; standardized; one SD = 1.6 points on the original 0-8 scale)	0.001	1.001	≈ no meaningful association

Looking at the table, the ordinal logistic regression coefficients are reported in log-odds units called odds ratios (ORs), which are obtained by exponentiating the coefficients and are presented for ease of interpretation. They directly reflect the change in odds of being in a higher alcohol-risk category, not movement between specific adjacent categories, as the model does not assume equal spacing between the categories. The percentage change in odds in the "Brief In-

terpretation" value is calculated by subtracting 1 from the odds ratio and multiplying the value by 100. The two variables with the highest impact on the model were RSBQR and N3Q77A. Among otherwise similar students, those who were screened positive for suicidal behavior have 20% higher odds of being in a higher alcohol-risk category than a lower category. Similarly, according to the model, students who are involved in Greek life have 20% higher odds of being in a higher alcohol-risk category than any lower category. In addition, the model indicates that students who have SPD have 10% greater odds of higher risk of alcohol misuse, and students who are lonely have 9% greater odds. For each increase in GPA values (B- to B, or B to B+, for example), the odds of being at greater risk of alcohol misuse are 8% lower. Some other interesting findings from the other variables were the complete lack of significant effect from variables such as insomnia, sleep apnea, and OCD diagnoses, as well as scores on the resilience scale. Additionally, according to the model, there are 4x greater odds for higher alcohol risk when a student is a Greek organization member compared to when they live in Greek housing, which is contrary to what was found in the EDA, as students who lived in Greek housing seemed to be at greater risk than those who simply participate in a Greek organization.

The next section of this report will further cover the results of the modeling process and connect the findings to those from the review of the literature.

4 Results and Discussion

In this study, data cleaning and analysis were performed using Python modules such as NumPy, Pandas, Scikit-Learn, and Statsmodels. During this phase, the sample size of the data was 103,639 students at its highest, and many correlations were drawn between alcohol risk and predictive variables. These include demographics such as sex, year in college, participation in Greek Life, Greek living situation, and student GPA and health-related factors such as the status of mental and behavioral health condition diagnoses and other variables that summarize a student's levels of loneliness, suicidal behavior, serious psychological stress, resilience, and overall well-being. The most prominent takeaways from the analysis and visualization phase were the connections between risk of alcohol misuse and Greek living situation, GPA, the presence of mental health conditions, and the summarized mental health scores. In addition, relationships between mental health conditions and scores and more specific questions on alcohol were investigated, such as problems a student has encountered due to alcohol use, failed attempts to control, reduce or stop alcohol use, and undesirable behavioral factors such as experiencing a blackout or driving under the influence. These relationships were further investigated during the modeling process.

After data collection, cleaning, analysis, and visualization, an ordinal regression model was developed to further analyze the relationships between the risk of alcohol use and the aforementioned predictive factors. Some additional cleaning was necessary before testing the model, which reduced the sample size to 58,933 students. 75% of this group was used to train the model, and the other 25% was used to test the model. Testing the model resulted in a MAE score of 0.278, indicating that, on average, the model's predictions were less than 0.3 categories off, where the jump from low to moderate risk, or moderate to high risk, is one category. In addition to the strong predictive power indicated by the MAE, the model also predicted 98.7% of cases either exactly how they were or only by one category, meaning that only 1.3% of the test data predicted a low risk case when the actual was high risk, and vice versa.

Furthermore, the predictive impact on the model from each variable was shown. The strongest predictors on the model were shown when a student had suicidal behavior (OR = 1.204) and/or

was a member of a fraternity or sorority (OR = 1.199), where the presence of both of these factors each increased the odds of alcohol misuse risk by 20%. In previous studies, rates of abusive drinking were notably higher for students involved in traditional college folklore, including Greek life hazing, initiation, and general participation [9]. Additionally, hazardous alcohol consumption was more likely to occur when the student's circle of friends drank often and when the student has mental health issues [5]. Greek life participation has also been shown to correlate positively with all but one of the alcohol-related questions on the AUDIT, used to measure alcohol use disorders [22]. Students who were positive for psychological distress (OR = 1.096) or loneliness (1.087) also experienced some of the highest odds of greater alcohol misuse risk in this model, at 10% and 9%, respectively. In previous literature, students with weaker mental health indicators experienced more alcohol-related issues than those without [12].

In this model, as a student's GPA increased (based on letter grade: B- to B, B to B+, etc.), the chances of that student having greater alcohol risk were 8% less (OR = 0.919). Previous studies have shown that academic issues like missing classes and receiving poor grades are more likely to arise as a result of excessive drinking [4] [17] [18].

While, in this model, each increase in perceived stress level only increased the odds of alcohol misuse by 4% (OR = 1.044), several other previous studies have identified stress as being rather impactful. University students who were high-risk drinkers were often those with higher stress levels, and many of these students coped with their stress using avoidance mechanisms [6].

Similarly to stress, the impact of anxiety on the model was subtle in comparison to many other predictors, where students with anxiety had 3% lower odds of greater alcohol misuse (OR = 0.970). The existence of depression in students had moderately stronger effects of increasing the odds of alcohol misuse at 7% (OR = 1.068). In previous studies, students with elevated depressed mood or anxiety symptoms were shown to be especially at risk when they hold higher descriptive perceptions, meaning that depressed or anxious students who believe their peers drink more frequently or in larger quantities are more likely to experience alcohol-related issues. These students are also more likely to use alcohol to cope with their issues. Additionally, these consequences were more sparse among students with stronger mental health. [12]. The impacts of anxiety and depression on predicting alcohol misuse risk were weak-moderate, which is a little bit less impactful than what is typically seen in the relevant literature. Perhaps this may be due to some students having been "diagnosed" for a while, but having found ways to cope or having acute symptoms, whereas the mental health scores were gathered from questions answered at the time of the survey. Current alcohol risk is tied to current symptoms of mental health issues.

In terms of other mental health issues, the presence of insomnia (OR = 0.994), sleep apnea (OR = 0.997), OCD (OR = 0.997), and lower levels of resilience (OR = 1.001) had no meaningful impact on the model. In contrast to these model results, OCD was previously shown to have high comorbidity with substance use risk and disorder, where the presence of OCD was associated with elevated odds of medium to high alcohol use risk among college students [23]. The presence of ADHD had a slightly greater effect (OR = 1.045), increasing the odds of greater alcohol risk by 4%. Previous correlations have been investigated between the presence of ADHD and alcohol use, where ADHD was associated with higher AUDIT scores, especially in categories regarding hazardous behavior and dependency on alcohol. [22]. Additionally, student well-being was negatively associated with higher alcohol misuse risk (OR = 0.924), where each one–standard deviation increase in the DIENER scale (approximately an 8.6-point increase on the original 8–56 scale) was associated with an 8% reduction in the odds of being in a higher alcohol risk category. Gambling disorder, while showing strong association with alcohol misuse risk in the EDA section, was excluded from the modeling process due to a very small sample size,

comparatively. While the majority of the literature supports the rational idea that mental health issues and diagnoses can put an individual at higher risk of alcohol misuse, this study sheds light on the idea that mental health issues themselves may be more impactful than the diagnoses themselves in predicting alcohol misuse risk.

Overall, the modeling process gave deeper insights into the relationships investigated in the EDA. The ordinal regression model revealed some relationships that were directly paralleled in the literature review, such as the connections between alcohol misuse risk, as well as other drinking habits, and Greek life participation, general mental health issues, and academic performance. Some relationships were noticeably different and weaker than expected, however, those being the majority of mental health diagnoses as well as the student's resilience scores. The presence of depression, ADHD, and anxiety each showed moderate impact on predicting risk of alcohol misuse, while PTSD, insomnia, sleep apnea, and OCD, showed minimal impact. The final main section of this report will summarize the overall findings and implications of this study and highlight their relevance to existing research and their foundation for future research.

5 Conclusion

This study began as a general investigation of alcohol-related issues and consequences among college students. However, through the methodology process, the research evolved into an effort of capturing ordinal relationships between predictors and alcohol misuse risk. Using the ACHA-NCHA III spring 2024 dataset, general statistics on alcohol-related issues and consequences were assembled, and EDA and visualization were conducted to investigate relationships prior to modeling. The ordinal logistic regression model was created after having reviewed a variety of topical literature, completing the EDA and visualization process, and checking for any collinearity issues through correlation matrices and VIF. The model performed well at distinguishing between increasing levels in alcohol misuse risk, demonstrating meaningful associations between incorporated predictors and the outcome variable with solid performance. However, due to class-size imbalance, estimates for the higher-risk categories, which are more sparsely represented, should be interpreted with caution, as these estimate are less precise and indicate limited stability and generalizability for these higher-risk groups. The MAE score of 0.278 indicated that, on average, the model's predictions were less than 0.3 categories off, where the jump from low to moderate risk, or moderate to high risk, is one category. Additionally, the model also predicted 98.7% of cases exactly how they were or one category off.

Key findings indicated that the strongest predictors of higher odds of alcohol misuse were suicidal behavior (OR = 1.204, indicating approximately 20% higher odds) and membership in a Greek organization (fraternity/sorority; OR = 1.199, approximately 20% higher odds). Additionally, students living in Greek housing had about 5% higher odds of alcohol misuse risk (OR = 1.051). This data indicates that students who participate in Greek life, as well as those who live in Greek housing, need to be educated on aware of the numerous possible consequences of alcohol misuse, as these students are at a statistically higher risk than the average, non-Greek-participating student. This sentiment is echoed by the majority of sources in the literature that investigated this relationship as well.

Other general mental health issues such as psychological distress (OR = 1.096, indicating 10% higher odds), loneliness (OR = 1.087, indicating 9% higher odds) showed moderately strong predictive power, and overall-well being (OR = 0.924, indicating 8% higher odds as quality of life decreases). However, mental health disorder diagnoses themselves showed weaker predictive power. Depression was the strongest predictor of these (OR = 1.068, indicating 7% higher odds),

followed by ADHD (OR = 1.045, indicating 4% higher odds), anxiety (OR = 0.970, indicating 3% lower odds), PTSD (OR = 1.012, indicating 1% higher odds), and insomnia, sleep apnea, and OCD all with no meaningful association to alcohol misuse risk. This indicates that the presence of mental health disorder symptoms may be more informative at predicting alcohol misuse risk than formal mental health disorder diagnoses themselves. Symptoms such as suicidal behavior, psychological distress, and loneliness likely reflect students' current emotional status, which may more directly influence alcohol use behaviors. In contrast, mental health diagnoses are more broad, and may include whose symptoms are now well managed and no longer acute. The findings from this study suggest that screening for symptoms of mental health issues may be more important than relying on diagnosed conditions in identifying students who are at elevated risk of alcohol misuse.

Additionally, in the ordinal regression model, each increase in a student's perceived stress level (no stress/low/moderate/high) raised the odds of greater alcohol risk by 4% (OR = 1.044), indicating that stress has mild-to-moderate impact on the risk of alcohol misuse in college students. There are two other predictors that were discussed in passing, those being each increase a student's overall health rating and each increase in the standardized CDRISC2. The model indicated that, for each increase in overall health rating (poor/fair/good/very good/excellent), a student's odds of greater alcohol risk increase by 2% (OR = 1.024), which is opposite to many of the related predictors, the majority of which have higher impacts on the model. Although both collinearity checks indicate balance across the spread of predictors, perhaps this variable still endured some overlap with others. CDRISC2, a student's resilience score, standardized for modeling, showed no meaningful power in predicting alcohol misuse risk.

This implications of this study's findings reach beyond just college students, as institutions, those in administrative positions at colleges, and researchers may find this research to be informative and compelling. This study allows college students to learn about rates of alcohol-related consequences, reflect on their own drinking behaviors, and gain a deeper understanding of what factors contribute to higher risk of alcohol misuse. For institutions, these findings can help inform the development of targeted prevention strategies and education programs focused on high-risk groups. Similarly, for those in administrative positions at colleges, these findings bring attention to the importance of symptom-based screening, staff training, and prevention efforts, to recognize risk indicators. Finally, for researchers, this findings suggest that symptom-level mental health measures may be more predictive of alcohol misuse risk than mental health diagnoses themselves, and encourage future research utilizing mixed methods and introducing new predictors to form a deeper understanding of the connection between alcohol-related harm and the elements that cause it to happen.

Every study does have limitations, however. This study utilizes the ACHA-NCHA data, which is cross-sectional, and focused on the general health of college students and the multitude of factors that contribute to that, rather than just on alcohol. The majority of this data is self-reported as well, meaning that there is a lack of causal inference to be made here. However, because of the large sample size, and spread across hundreds of higher-education institutions, the correlative findings do hold moderate weight. Another weakness of this data is that there are inconsistent time frames between different questions, such as some questions being the last 2 weeks, some being the last 3 months, and some being the last year, is a little frustrating to work with, but with the data being self-reported, it cannot be expected that people will report accurately on everything if they are always asked about their actions in the last year, for example. To obtain findings on this topic that are optimally accurate and can causally be most confidently reported, a multitude of factors would need to be held constant and a control group would need to be present as well. This data absolutely gives great insights into trends among American college students, but cannot accurately report on causality.

For future studies, these factors should be investigated through different methods, and other predictors should be introduced as well. Studying what specific alcohol-related issues are the most harmful could provide more insight into the topic as well. As mentioned above, although carrying out a study with strict control parameters would allow findings to be reported with greater confidence, due to the nature of alcohol, and the large list of factors that would need to be accounted for, being able to create a study such as that would be very difficult.

Overall, this study reinforces the impact that mental health factors and participation in college Greek life have on alcohol misuse risk. This study also reveals the importance of approaches and interventions to inform people of alcohol-related issues and consequences and prevent these issues from arising.

6 Acknowledgments

First of all, I would like to thank my parents. They have always supported me and pushed me to strive for excellence, and so, without their support, I believe I truly would not be where I am today.

I would like to thank Adrian College for providing quality learning opportunities to students, such as allowing us to undertake projects like summer research and our capstone classes, where we are able to investigate topics interesting to us and practice our skills in reporting them.

I would also like to thank my research supervisor, Dr. Yasser Alginahi. I have had Dr. Alginahi as both a professor and research supervisor multiple times at this point, and he has been a great mentor to me in growing my research and analysis skills. He has always pushed me to get out of comfort zone in pursuit of knowledge, and this is a trait I believe will carry with me long after Adrian College.

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.

References

[1] MedlinePlus, U.S. National Library of Medicine, "Alcohol," https://medlineplus.gov/alcohol.html, 2024, accessed: 2025-11-16.

[2] National Highway Traffic Safety Administration, "Drunk driving — statistics and resources," https://www.nhtsa.gov/risky-driving/drunk-driving, 2023, accessed: 2025-11-16.

[3] National Institute on Alcohol Abuse and Alcoholism, "Alcohol and young adults ages 18 to 25," https://www.niaaa.nih.gov/alcohols-effects-health/alcohol-topics/alcohol-facts-and-statistics/alcohol-and-young-adults-ages-18-25, 2025, accessed: 2025-11-16.

[4] A. White and R. Hingson, "The burden of alcohol use: Excessive alcohol consumption and related consequences among college students," *Alcohol Research: Current Reviews*, vol. 35, no. 2, p. 201, 2014.

[5] A. Ay, C. Çam, A. Kılınç, M. F. Önsüz, and S. Metintaş, "Prevalence of hazardous alcohol consumption and evaluation of associated factors in university students," *Social Psychiatry and Psychiatric Epidemiology*, vol. 60, no. 2, pp. 223–233, 2025.

[6] M. S. C. Chow, S. H. L. Poon, K. L. Lui, C. C. Y. Chan, and W. W. T. Lam, "Alcohol consumption and depression among university students and their perception of alcohol use," *East Asian Archives of Psychiatry*, vol. 31, no. 4, pp. 87–96, 2021.

[7] American College Health Association, "National college health assessment iii: Reference group executive summary spring 2022," Hanover, MD, 2022, [Online]. Available: https://www.unthsc.edu/care-and-civility/national-college-health-assessment-iii-2022.

[8] D. C. R. Kerr, M. R. Kasimanickam, D. E. Bradford, H. Bae, and K. A. Parks, "Problematic alcohol behaviors and sexual assault on college campuses: How are student reports and institution-reported crime data related?" *Journal of American College Health*, 2025.

[9] V. Lorant, P. Nicaise, V. E. Soto, and W. d'Hoore, "Alcohol drinking among college students: college responsibility for personal troubles," *BMC Public Health*, vol. 13, pp. 1–9, 2013.

[10] M. Herrero-Montes, C. Alonso-Blanco, M. Paz-Zulueta, A. Pellico-López, L. Ruiz-Azcona, C. Sarabia-Cobo, and P. Parás-Bravo, "Excessive alcohol consumption and binge drinking in college students," *PeerJ*, vol. 10, p. e13368, 2022.

[11] M. P. Davoren, J. Demant, F. Shiely, and I. J. Perry, "Alcohol consumption among university students in ireland and the united kingdom from 2002 to 2014: a systematic review," *BMC Public Health*, vol. 16, no. 1, p. 173, 2016.

[12] S. R. Kenney, G. T. DiGuiseppi, M. K. Meisel, S. G. Balestrieri, and N. P. Barnett, "Poor mental health, peer drinking norms, and alcohol risk in a social network of first-year college students," *Addictive Behaviors*, vol. 84, pp. 151–159, 2018.

[13] M. R. Pearson, A. M. Bravo, and M. M. Conner, "A quantification of the alcohol use–consequences association in college student and clinical populations: A large, multi-sample study," *Psychology of Addictive Behaviors*, vol. 32, no. 3, pp. 266–278, 2018.

[14] G. W. Shorter, P. Leonard, B. Bunting, J. Skelly, N. Miller, and C. Campbell, "Alcohol consumption and attitudes to evidence-based alcohol policy in donegal: Findings from a student and general adult sample," 2022, [Online]. Available: https://pureadmin.qub.ac.uk/ws/portalfiles/portal/322236162/Donegal_student_and_population_report_final_v2.pdf.

[15] A. Lukács, N. Simon, J. S. Demeter, É. Kissné Dányi, and E. Kiss-Tóth, "Alcohol consumption among university students," *Egészségtudományi közlemények: a miskolci egyetem közleménye*, vol. 3, no. 2, pp. 57–61, 2013.

[16] Y. Tang, C. G. Abildso, C. L. Lilly, E. L. Winstanley, and T. M. Rudisill, "Risk factors associated with driving after marijuana use among u.s. college students during the covid-19 pandemic," *Journal of Adolescent Health*, vol. 72, no. 4, pp. 544–552, 2023, [Online]. Available: https://pmc.ncbi.nlm.nih.gov/articles/PMC9637518/pdf/main.pdf.

[17] National Institute on Alcohol Abuse and Alcoholism, "College drinking fact sheet," Oct. 2020, [Online]. Available: https://www.campusdrugprevention.gov/sites/default/files/2021-11/NIAAA_CollegeDrinking_Oct2020.pdf.

[18] ——, "Fall semester—a time for parents to discuss the risks of college drinking," 2023, [Online]. Available: https://www.prnewswire.com/news-releases/niaaa-fall-semestera-time-for-parents-to-discuss-the-risks-of-college-drinking-302215446.html.

[19] K. Yoder, D. Dzemidzic, D. Kareken *et al.*, "Differences in iv alcohol-induced dopamine release in the ventral striatum of social drinkers and nontreatment-seeking alcoholics," *Drug and Alcohol Dependence*, vol. 158, pp. 188–195, 2016.

[20] D. Conroy and R. de Visser, "Benefits and drawbacks of social non-drinking identified by british university students," *Drug and Alcohol Review*, vol. 36, no. 2, pp. 251–259, 2017, [Online]. Available: https://eprints.bbk.ac.uk/id/eprint/22615/1/22615.pdf.

[21] R. de Visser, E. Robinson, and R. Bond, "Voluntary temporary abstinence from alcohol during 'dry january' and subsequent alcohol use," *Psychology & Health*, vol. 35, no. 3, pp. 359–374, 2020, [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/26690637/.

[22] M. Rooney, A. Chronis-Tuscano, and Y. Yoon, "Substance use in college students with adhd," *Journal of attention disorders*, vol. 16, no. 3, pp. 221–234, 2012.

[23] W. Jacobs, A. DeLeon, A. Bristow, P. Quinn, and A. Lederer, "Substance use and disordered eating risk among college students with obsessive-compulsive conditions," *PLoS One*, vol. 20, no. 1, p. e0316349, 2025.

Appendices

6.1 Appendix A: Cleaning

```
'''
Jonah Watson
Spring 2026
Mental and Behavioral Consequences of Alcohol Consumption Among College Students
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
import json             # imported to access the JSON file that stores column label mappings



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

# open the original CSV file to access the text responses for N3Q28 and fix a mapping issue
df28 = pd.read_csv("NCHA-III S24 - Labeled.csv", usecols=["N3Q28"], keep_default_na=False)

# load the JSON label mappings
with open("NCHA-IIIb labels S24 - Cleaned.json", "r") as f:
    labels = json.load(f)

mapping_28 = labels["N3Q28"]
# reverse mapping: text to numeric
reverse_map = {v.strip().lower(): int(float(k)) for k, v in mapping_28.items()}

# this function handles missing inputs and maps text responses to numeric codes
def map_n3q28(val):
    val_str = str(val).strip().lower()      # normalize the input value
    if val_str == "":                       # check for empty strings
        return pd.NA
    return reverse_map.get(val_str, pd.NA)
    # look for standardized value in the reverse mapping dictionary and return pd.NA if not found

# apply the mapping function to the N3Q28 column
df["N3Q28"] = df28["N3Q28"].apply(map_n3q28)

# keep specific columns (removing unnecessary variables from the dataset)
keep_variables = ["N3Q1", "N3Q22B2", "N3Q22K2","N3Q22L2",
    "N3Q22M2","N3Q22N2","N3Q22O2","N3Q25B1","N3Q25B2","N3Q28", "N3Q29A","N3Q29B",
    "N3Q29C","N3Q29D","N3Q29E","N3Q29F","N3Q29G","N3Q29H","N3Q29I","N3Q29J","N3Q29K", "N3Q29L",
    "N3Q30A", "N3Q30B", "N3Q46", "N3Q48", "N3Q65A2","N3Q65A3","N3Q65A7", "N3Q65A15", "N3Q65A19",
    "N3Q65A28","N3Q65A31", "N3Q65A33","N3Q65A35","N3Q65Y","N3Q67A", "N3Q69", "N3Q72", 'N3Q75A1',
    'N3Q75A2', 'N3Q75A3', 'N3Q75A4', 'N3Q75A5', 'N3Q75A6', 'N3Q75A7', 'N3Q75A8', "N3Q77A",
    "N3Q77B", "N3Q80", "RKESSLER6","RULS3","RSBQR","DIENER", "CDRISC2", "ALCOHOLRISK"]

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

# save the cleaned numeric CSV
df_cleaned.to_csv("CLEANED ALCOHOL NCHA-III S24 - New_Numeric.csv", index=False)

# check that 'None' responses are now coded as 1 - should be 24530
print(df_cleaned["N3Q28"].value_counts(dropna=False))
print('Cleaned data saved')

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

6.2 Appendix B: Profiling

```
'''
Jonah Watson
Spring 2026
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 that stores column label mappings



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

# load the JSON label mappings
with open("NCHA-IIIb_labels_S24_copycopy.json", 'r') as jsonfile:
    labels = json.load(jsonfile)

# prepare a text file to store profiling data and response rates of each choice for each question
response_rates = "alcohol_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 my columns
        # 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 the name "Missing" in the text file
                # if the data exists, look at the JSON for what the numeric correlation is
                # example. in some questions "1" = "Excellent", "2" = "Very Good", 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
'''
```

6.3 Appendix C: Analyzing

```
'''
Jonah Watson
Spring 2026
Mental and Behavioral Consequences of Alcohol Consumption Among College Students
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 matplotlib.pyplot as plt                     # imported for data visualization
import numpy as np                                  # imported to perform operations on arrays
from sklearn.preprocessing import StandardScaler    # imported to standardize the data
from sklearn.decomposition import PCA               # imported to perform Principal Component Analysis (PCA)



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

# recode this question so that higher values indicate better overall health
# to be consistent for the format of other questions
df['N3Q1'] = 6 - df['N3Q1']    # 1 becomes 5, 2 becomes 4, etc.


# shortened lists of variables for analysis:

# alcohol variables
# alcohol_vars = ["N3Q22B2", "N3Q28", "ALCOHOLRISK"]
alcohol_vars = ["N3Q22L2", "N3Q22O2", "N3Q29B", "N3Q30A"]

# demographic variables
demographic_vars = ["N3Q67A", "N3Q72", "N3Q77A", "N3Q77B", "N3Q80"]

# health variables
health_vars = ["N3Q1", "N3Q48", "N3Q65A2", "N3Q65A7", "N3Q65A15",
"N3Q65A19", "N3Q65A28","N3Q65A31",
"N3Q65A33","N3Q65A35", "RKESSLER6","RULS3","RSBQR","DIENER", "CDRISC2"]



# prepare a text file to store correlations between alcohol and demographic variables
alc_demographic_file = "alc_demographic_correlations.txt"

"""The following block of code underneath the "with" statement will be reused and
slightly modified again for the health variable class that was created above as well."""
with open(alc_demographic_file, 'w') as f:
    for demographic in demographic_vars:
```

```
        for alc in alcohol_vars:

            # write the percentages within each demographic
            percentages = pd.crosstab(df[alc], df[demographic], normalize='columns') * 100
            f.write(f"\n{alc} 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")

print(f"Correlation analysis saved to {alc_demographic_file}\n\n")



# prepare a text file to store the correlations between alcohol and health variables
alc_health_file = "alc_health_correlations.txt"

with open(alc_health_file, 'w') as f:
    for health in health_vars:
        for alc in alcohol_vars:

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

print(f"Correlation analysis saved to {alc_health_file}\n\n")


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

# use a stacked bar chart to visualize the relationship between risk of alcohol misuse
# and the PRESENCE of specific disorders
disorder_labels = ['ADD\ADHD\n(n = 14,032)', 'anxiety\n(n = 35,840)', 'depression\n(n = 27,210)',
        'gambling disorder\n(n = 162)', 'insomnia\n(n = 7,239)', 'OCD\n(n = 6,133)',
        'PTSD\n(n = 8,292)', 'sleep apnea\n(n = 2,394)']
alc_risks = {
    "Low Risk": np.array([79.66, 81.79, 80.27, 48.91, 80.06, 79.71, 79.39, 81.78]),
    "Moderate Risk":   np.array([17.82, 16.33, 17.44, 29.35, 16.94, 17.45, 17.48, 14.92]),
    "High Risk":   np.array([2.51, 1.87, 2.29, 21.74, 3.00, 2.84, 3.13, 3.31])
}

width = 0.6

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

colors = {
    "Low Risk": "#4CAF50",      # green
    "Moderate Risk": "#FFC107", # yellow/orange
    "High Risk": "#F44336"      # red
}

for timerange, percent in alc_risks.items():
    ax.bar(disorder_labels, percent, width, label=timerange, bottom=bottom,
           color=colors[timerange], edgecolor='black', linewidth=0.6
    )
    bottom = bottom + percent

ax.set_title('Risk of Alcohol Misuse by PRESENCE of Disorders', fontsize=14, weight='bold')
ax.set_ylabel('Percentage of Students (%)', fontsize=12)
ax.set_xlabel('Disorders', fontsize=12)
ax.set_ylim(0, 100)
ax.legend(title='Risk of Alcohol Misuse', loc='upper left', bbox_to_anchor=(1.05, 1))
plt.xticks(rotation=45)
plt.tight_layout()
plt.savefig('ALCOHOLRISK_by_disorderPRESENCE.png')
```

```
# use a stacked bar chart to visualize the relationship between risk of alcohol misuse
# and the ABSENCE of specific disorders
disorder_labels = ['no ADD\ADHD\n(n = 87,472)', 'no anxiety\n(n = 65,895)',
    'no depression\n(n = 74,446)', 'no gambling disorder\n(n = 101,144)',
    'no insomnia\n(n = 93,560)', 'no OCD\n(n = 95,464)',
    'no PTSD\n(n = 93,279)', 'no sleep apnea\n(n = 98,852)']
alc_risks = {
    "Low Risk":        np.array([84.91, 85.58, 85.82, 84.17, 84.47, 84.42, 84.59, 84.20]),
    "Moderate Risk":   np.array([13.88, 13.31, 13.16, 14.46, 14.27, 14.28, 14.18, 14.45]),
    "High Risk":       np.array([1.21, 1.11, 2.29, 1.37, 1.27, 1.30, 1.22, 1.35])
}

width = 0.6

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

colors = {
    "Low Risk": "#4CAF50",      # green
    "Moderate Risk": "#FFC107", # yellow/orange
    "High Risk": "#F44336"      # red
}

for timerange, percent in alc_risks.items():
    ax.bar(disorder_labels, percent, width, label=timerange, bottom=bottom,
           color=colors[timerange], edgecolor='black', linewidth=0.6
    )
    bottom = bottom + percent

ax.set_title('Risk of Alcohol Misuse by ABSENCE of Disorders', fontsize=14, weight='bold')
ax.set_ylabel('Percentage of Students (%)', fontsize=12)
ax.set_xlabel('Disorders', fontsize=12)
ax.set_ylim(0, 100)
ax.legend(title='Risk of Alcohol Misuse', loc='upper left', bbox_to_anchor=(1.05, 1))
plt.xticks(rotation=45)
plt.tight_layout()
plt.savefig('ALCOHOLRISK_by_disorderABSENCE.png')



# use a stacked bar chart to visualize the relationship between risk of alcohol misuse and GPA
gpa_labels = [
    'A+', 'A', 'A-', 'B+', 'B', 'B-',
    'C+', 'C', 'C-', 'D+', 'D', 'D-', 'F'
]

alc_risks = {
    "Low Risk": np.array([
        86.96, 86.16, 83.47, 82.18, 82.23, 80.36,
        80.20, 81.81, 78.51, 69.68, 73.50, 73.68, 72.22
    ]),
    "Moderate Risk": np.array([
        11.80, 12.82, 15.23, 16.32, 15.97, 17.23,
        17.71, 15.92, 19.74, 25.81, 22.22, 18.42, 12.96
    ]),
    "High Risk": np.array([
        1.23, 1.01, 1.31, 1.51, 1.80, 2.41,
        2.09, 2.27, 1.75, 4.52, 4.27, 7.89, 14.81
    ])
}

width = 0.7
fig, ax = plt.subplots(figsize=(9, 6))
bottom = np.zeros(len(gpa_labels))


colors = {
    "Low Risk": "#4CAF50",      # green
    "Moderate Risk": "#FFC107", # yellow/orange
    "High Risk": "#F44336"      # red
}

for risk, percent in alc_risks.items():
```

```
    ax.bar(gpa_labels, percent, width, label=risk, bottom=bottom,
        color=colors[risk], edgecolor='black', linewidth=0.6
    )
    bottom = bottom + percent

# Formatting
ax.set_title('Risk of Alcohol Misuse by GPA', fontsize=14, weight='bold')
ax.set_ylabel('Percentage of Students (%)', fontsize=12)
ax.set_xlabel('GPA Category', fontsize=12)
ax.set_ylim(0, 100)
ax.legend(title='Risk of Alcohol Misuse', loc='upper left', bbox_to_anchor=(1.02, 1))
plt.xticks(rotation=0)
plt.tight_layout()
plt.savefig('ALCOHOLRISK_by_GPA.png')



# use a stacked bar chart to visualize the relationship between risk of alcohol misuse
# and Greek Life participation
greek_labels = ['Not members of a fraternity/sorority', 'Members of a fraternity/sorority']
alc_risks = {
    "Low Risk":        np.array([84.94, 73.69]),
    "Moderate Risk":   np.array([13.70, 24.16]),
    "High Risk":       np.array([1.36, 2.15])
}

width = 0.6

fig, ax = plt.subplots(figsize=(8,8))
bottom = np.zeros(len(greek_labels))

colors = {
    "Low Risk": "#4CAF50",      # green
    "Moderate Risk": "#FFC107", # yellow/orange
    "High Risk": "#F44336"      # red
}

for timerange, percent in alc_risks.items():
    ax.bar(greek_labels, percent, width, label=timerange, bottom=bottom,
           color=colors[timerange], edgecolor='black', linewidth=0.6
    )
    bottom = bottom + percent

ax.set_title('Risk of Alcohol Misuse by Greek Life Participation', fontsize=14, weight='bold')
ax.set_ylabel('Percentage of Students (%)', fontsize=12)
ax.set_xlabel('Greek Life Participation', fontsize=12)
ax.set_ylim(0, 100)
ax.legend(title='Risk of Alcohol Misuse', loc='upper left', bbox_to_anchor=(1.05, 1))
plt.xticks(rotation=45)
plt.tight_layout()
plt.savefig('ALCOHOLRISK_by_greek_life_particpation.png')



# use a stacked bar chart to visualize the relationship between frequency of alcohol use
# in the last 3 months and Greek Life living
greek_labels = ['Not living in a fraternity/sorority', 'Living in a fraternity/sorority']
alc_rates = {
    "Never":           np.array([8.74, 3.38]),
    "Once or twice":   np.array([31.37, 13.04]),
    "Monthly":         np.array([26.06, 18.51]),
    "Weekly":          np.array([31.43, 60.54]),
    "Daily/almost":    np.array([2.40, 4.54])
}

width = 0.6

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

colors = {
    "Never": "#4CAF50",         # green
    "Once or twice": "#8BC34A", # light green
```

```
    "Monthly": "#FFC107",       # yellow/orange
    "Weekly": "#FF9800",        # orange
    "Daily/almost": "#F44336"   # red
}

for timerange, percent in alc_rates.items():
    ax.bar(greek_labels, percent, width, label=timerange, bottom=bottom,
           color=colors[timerange], edgecolor='black', linewidth=0.6
    )
    bottom = bottom + percent

ax.set_title('Frequency of Alcohol Consumption by Greek Life Living Situation',
             fontsize=14, weight='bold')
ax.set_ylabel('Percentage of Students (%)', fontsize=12)
ax.set_xlabel('Greek Life Living Situation', fontsize=12)
ax.set_ylim(0, 100)
ax.legend(title='Frequency of Alcohol Consumption\n(last 3 months)',
          loc='upper left', bbox_to_anchor=(1.05, 1))
plt.tight_layout()
plt.savefig('alcohol_consumption_by_greek_life_living.png')



# use a stacked bar chart to visualize the relationship between risk of alcohol misuse
# and Greek Life living
greek_labels = ['Not living in a fraternity/sorority', 'Living in a fraternity/sorority']
alc_risks = {
    "Low Risk":        np.array([84.27, 67.36]),
    "Moderate Risk":   np.array([14.35, 27.85]),
    "High Risk":       np.array([1.38, 4.79])
}

width = 0.6

fig, ax = plt.subplots(figsize=(9,6))
bottom = np.zeros(len(greek_labels))

colors = {
    "Low Risk": "#4CAF50",      # green
    "Moderate Risk": "#FFC107", # yellow/orange
    "High Risk": "#F44336"      # red
}

for timerange, percent in alc_risks.items():
    ax.bar(greek_labels, percent, width, label=timerange, bottom=bottom,
           color=colors[timerange], edgecolor='black', linewidth=0.6
    )
    bottom = bottom + percent

ax.set_title('Risk of Alcohol Misuse by Greek Life Living Situation',
             fontsize=14, weight='bold')
ax.set_ylabel('Percentage of Students (%)', fontsize=12)
ax.set_xlabel('Greek Life Living Situation', fontsize=12)
ax.set_ylim(0, 100)
ax.legend(title='Risk of Alcohol Misuse', loc='upper left', bbox_to_anchor=(1.05, 1))
#plt.xticks(rotation=45)
plt.tight_layout()
plt.savefig('ALCOHOLRISK_by_greek_life_living.png')



# use a stacked area chart to visualize the relationship between DIENER scores
# and risk of alcohol misuse
diener_scores = np.arange(8, 57)

low = np.array([
    73.73,76.67,78.79,67.86,82.76,75.68,70.97,74.00,57.45,74.65,
    65.43,73.58,75.23,67.74,72.00,73.25,71.01,73.36,76.64,75.32,
    77.15,74.20,75.39,74.75,77.97,80.68,80.00,81.28,79.94,81.12,
    79.52,83.19,81.01,81.50,82.59,83.36,83.97,83.15,84.58,84.14,
    86.46,86.11,86.18,87.21,87.02,86.48,86.13,87.03,88.78
])
```

```
moderate = np.array([
    18.64,16.67,18.18,21.43,10.34,21.62,22.58,18.00,35.11,19.72,
    29.63,23.58,18.35,27.42,20.67,24.20,22.69,22.71,18.61,21.47,
    18.18,20.88,20.51,22.20,17.07,16.56,16.61,16.72,17.60,16.61,
    18.88,15.18,16.86,17.05,15.83,15.23,14.68,15.82,14.35,14.91,
    12.53,13.18,13.01,12.12,12.46,13.03,13.16,12.61,10.44
])

high = np.array([
    7.63,6.67,3.03,10.71,6.90,2.70,6.45,8.00,7.45,5.63,
    4.94,2.83,6.42,4.84,7.33,2.55,6.30,3.93,4.74,3.21,
    4.67,4.91,4.10,3.05,4.96,2.76,3.39,2.00,2.47,2.26,
    1.60,1.63,2.13,1.45,1.58,1.42,1.35,1.03,1.07,0.94,
    1.01,0.71,0.82,0.66,0.52,0.50,0.71,0.37,0.79
])

# set up trend line
moderate_high = moderate + high

coef = np.polyfit(diener_scores, moderate_high, 1)
trend_line = np.poly1d(coef)(diener_scores)

colors = {
    "Low Risk": "#4CAF50",      # green
    "Moderate Risk": "#FFC107", # yellow/orange
    "High Risk": "#F44336"      # red
}

plt.figure(figsize=(11, 6))

plt.stackplot(diener_scores, high, moderate, low,
    labels=['High Risk', 'Moderate Risk', 'Low Risk'],
    colors=[colors["High Risk"], colors["Moderate Risk"], colors["Low Risk"]],
    alpha=0.85
)

# plot trend line
plt.plot(
    diener_scores,
    trend_line,
    color='black',
    linewidth=2.5,
    linestyle='--',
    label='Trend: Moderate{High Risk'
)

plt.title('Alcohol Misuse Risk Distribution by Well-Being (DIENER)',
          fontsize=14, weight='bold')
plt.xlabel('DIENER Flourishing Score', fontsize=12)
plt.ylabel('Percentage of Students (%)', fontsize=12)
plt.ylim(0, 100)
plt.legend(loc='upper right')
plt.grid(alpha=0.3)
plt.tight_layout()
plt.savefig('ALCOHOLRISK_by_DIENER_stacked_area.png', dpi=300)



# create a stacked bar chart to visualize the relationship between CDRISC2 scores
# and risk of alcohol misuse
cdrisc2_labels = ['0 (not resilient)', '1', '2', '3', '4 (somewhat resilient)',
                  '5', '6', '7', '8 (very resilient)']
alc_risks = {
    "Low Risk": np.array([68.15, 77.56, 75.82, 77.52, 80.70, 82.65, 84.87, 85.13, 86.35]),
    "Moderate Risk": np.array([22.18, 20.49, 19.48, 19.53, 16.92, 15.73, 13.07, 13.90, 12.68]),
    "High Risk":   np.array([9.68, 1.95, 4.71, 2.95, 2.38, 1.63, 1.17, 0.97, 0.98])
}

width = 0.6

fig, ax = plt.subplots(figsize=(10,6))
bottom = np.zeros(len(cdrisc2_labels))
```

```
colors = {
    "Low Risk": "#4CAF50",      # green
    "Moderate Risk": "#FFC107", # yellow/orange
    "High Risk": "#F44336"      # red
}

for timerange, percent in alc_risks.items():
    ax.bar(cdrisc2_labels, percent, width, label=timerange, bottom=bottom,
           color=colors[timerange], edgecolor='black', linewidth=0.6
    )
    bottom = bottom + percent

ax.set_title('Risk of Alcohol Misuse by CDRISC2 Score', fontsize=14, weight='bold')
ax.set_ylabel('Percentage of Students (%)', fontsize=12)
ax.set_xlabel('CDRISC2 Score (Higher Scores Indicate Greater Resilience)', fontsize=12)
ax.set_ylim(0, 100)
ax.legend(title='Risk of Alcohol Misuse', loc='upper left', bbox_to_anchor=(1.05, 1))
plt.xticks(rotation=45)
plt.tight_layout()
plt.savefig('ALCOHOLRISK_by_CDRISC2.png')



# create a stacked bar chart to visualize the relationship between alcohol-related outcomes
# and Greek Life living situation
questions = [
    'Health/social/legal/financial\nproblems from alcohol',
    'Failed to control\nalcohol use',
    'Blackout while\ndrinking',
    'Drove after\ndrinking'
]

not_greek = np.array([10.7, 8.4, 9.9, 12.8])
greek = np.array([23.8, 10.6, 26.7, 11.6])

x = np.arange(len(questions))
width = 0.6

fig, ax = plt.subplots(figsize=(10, 6))

colors = {
    "Not Greek": "#4CAF50",   # green
    "Greek": "#F44336"        # red
}

ax.bar(x, not_greek, width, label='Not in Greek housing',
    color=colors["Not Greek"], edgecolor='black', linewidth=0.6
)

ax.bar(x, greek, width, bottom=not_greek,
    label='In Greek housing', color=colors["Greek"], edgecolor='black', linewidth=0.6
)

ax.set_title('Alcohol-Related Outcomes by Greek Life Living Situation',
             fontsize=14, weight='bold')
ax.set_ylabel('Percentage of Students (%)', fontsize=12)
ax.set_xlabel('Alcohol-Related Outcomes', fontsize=12)
ax.set_xticks(x)
ax.set_xticklabels(questions, rotation=20, ha='right')
ax.set_ylim(0, 40)
ax.legend(title='Living Situation', loc='upper right')

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



# create a stacked bar chart to visualize the relationship between DIENER scores
# and presence of alcohol-related problems
diener_bins = [
    'Very low\n(8{17)',
    'Low\n(18{27)',
    'Moderate\n(28{37)',
```

```
    'High\n(38{47)',
    'Very high\n(48{56)'
]

# Percentages (example structure | plug in your computed values)
no_problems = np.array([80.3, 79.9, 84.4, 88.2, 91.4])
any_problems = np.array([19.7, 20.1, 15.6, 11.8, 8.6])

width = 0.7
x = np.arange(len(diener_bins))

fig, ax = plt.subplots(figsize=(10, 6))

colors = {
    "No problems": "#4CAF50",   # green
    "Any problems": "#F44336"   # red
}

ax.bar(x, no_problems, width, label='No alcohol-related problems',
    color=colors["No problems"], edgecolor='black', linewidth=0.6
)

ax.bar(x, any_problems, width,bottom=no_problems, label='Any alcohol-related problems',
    color=colors["Any problems"], edgecolor='black', linewidth=0.6
)

ax.set_title(
    'Any Alcohol-Related Problems by Well-Being (DIENER)', fontsize=14, weight='bold'
)

ax.set_ylabel('Percentage of Students (%)', fontsize=12)
ax.set_xlabel('DIENER Well-Being Score', fontsize=12)
ax.set_ylim(0, 100)
ax.set_xticks(x)
ax.set_xticklabels(diener_bins)

ax.legend(
    title='Alcohol-Related Problems',
    loc='upper left',
    bbox_to_anchor=(1.02, 1)
)

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

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

6.4 Appendix D: Modeling

```
'''
Jonah Watson
Spring 2026
Mental and Behavioral Consequences of Alcohol Consumption Among College Students
This file was used to test an ordinal logistic regression model predicting risk of alcohol misuse
'''


from sklearn.model_selection import train_test_split
# imported to split data into training and testing sets

from sklearn.preprocessing import StandardScaler
# imported to standardize continuous variables

from sklearn.pipeline import Pipeline
# imported to create modeling pipelines

from sklearn.metrics import accuracy_score, precision_score, recall_score, f1_score
# imported to evaluate model performance
```

```
from sklearn.metrics import classification_report, confusion_matrix, ConfusionMatrixDisplay
# imported to create and display confusion matrices

from statsmodels.stats.outliers_influence import variance_inflation_factor
# imported to check for multicollinearity among variables

import statsmodels.api as sm
# imported for statistical modeling

from statsmodels.miscmodels.ordinal_model import OrderedModel
# imported to perform ordinal logistic regression

import matplotlib.pyplot as plt   # imported for data visualization
import numpy as np                # imported for numerical operations
import pandas as pd               # imported for data manipulation and analysis
import seaborn as sns             # imported for enhanced data visualization



# ================
# Data Preparation
# ================

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

# performing answer recoding for consistency
binary_vars = [
    "N3Q65A2", "N3Q65A3", "N3Q65A7", "N3Q65A15",
    "N3Q65A19", "N3Q65A28", "N3Q65A31",
    "N3Q65A33", "N3Q65A35",
    "N3Q77A", "N3Q77B",
    "RULS3", "RSBQR"
]
# "N3Q25B1", "N3Q25B2", "N3Q29A", "N3Q29B", "N3Q29C", "N3Q29D", "N3Q29E",
# "N3Q29F", "N3Q29G", "N3Q29H", "N3Q29I", "N3Q29J", "N3Q29K", "N3Q29L", "N3Q30A",

for col in binary_vars:
    df[col] = df[col].replace({1: 0, 2: 1})

# standardizing continuous scale variables so they have mean=0 and std=1
scale_vars = [
    "DIENER",
    "CDRISC2"
]

scaler = StandardScaler()
df[scale_vars] = scaler.fit_transform(df[scale_vars])

# view how much each standard deviation unit corresponds to in original units
diener_sd = scaler.scale_[scale_vars.index("DIENER")]
print(f"DIENER standard deviation weight: {diener_sd}")
cdrisc2_sd = scaler.scale_[scale_vars.index("CDRISC2")]
print(f"CDRISC2 standard deviation weight: {cdrisc2_sd}")

# special case
df["RKESSLER6"] = df["RKESSLER6"].replace({1: 0, 3: 1})

# drop NA from variables that will be used in modeling
model_variables = binary_vars + scale_vars + ["N3Q1", "N3Q48", "N3Q80", "RKESSLER6", "ALCOHOLRISK"]
df = df.dropna(subset=model_variables)

# flip overall health so higher is better
df['N3Q1'] = 6 - df['N3Q1']

# flip GPA so higher is better
df["N3Q80"] = df["N3Q80"].max() - df["N3Q80"] + 1



# choose which variables will be included in the modeling process
X = df[["N3Q1",      # Overall health
```

```
    "N3Q48",         # Stress
    "N3Q77A",        # Greek membership
    "N3Q77B",        # Greek housing
    "N3Q80",         # GPA

    # mental health diagnoses
    "N3Q65A2",       # ADHD
    "N3Q65A7",       # Anxiety
    "N3Q65A15",      # Depression
    "N3Q65A28",      # Insomnia
    "N3Q65A31",      # OCD
    "N3Q65A33",      # PTSD
    "N3Q65A35",      # Sleep apnea

    # mental-Health Scales
    "RKESSLER6",     # Psychological distress scale
    "RULS3",         # Loneliness scale
    "RSBQR",         # Suicide risk scale
    "DIENER",        # Well-being scale
    "CDRISC2"        # Resilience scale
]]

# remap target variable
y = df["ALCOHOLRISK"].map({
    1: 0,
    2: 1,
    3: 2,
})



# ==============================
# COLLINEARITY CHECKS
'''check for collinearity among groups of variables to ensure no variables are
too highly correlated to be included together in the model'''
# ==============================



# mental health diagnoses
# the highest value was 0.65, which is good, as above 0.8 indicates high correlation
# which is undesirable for modeling
diagnoses = ["N3Q65A2", "N3Q65A3", "N3Q65A7", "N3Q65A15", "N3Q65A19", "N3Q65A28",
             "N3Q65A31", "N3Q65A33", "N3Q65A35"]
correlation_matrix = df[diagnoses].corr()
print(correlation_matrix)
plt.figure(figsize=(8,6))
sns.heatmap(correlation_matrix, annot=True, cmap="coolwarm", vmin=-1, vmax=1)
plt.title("Correlation Matrix of Mental Health Diagnoses")
plt.savefig("mental_health_diagnoses_correlation_matrix.png")

# Variance Influence Factor (VIF) calculation to ensure no multicollinearity among mental health scales
# the highest value was 2.65, which is good, as above 5 indicates high multicollinearity
# which is undesirable for modeling
X_diagnoses = df[diagnoses]
vif_data = pd.DataFrame()
vif_data["feature"] = X_diagnoses.columns
vif_data["VIF"] = [variance_inflation_factor(X_diagnoses.values, i) for i in range(X_diagnoses.shape[1])]
print(vif_data)



# mental health scales
# the highest value was 0.53, which is good, as above 0.8 indicates high correlation
# which is undesirable for modeling
scales = ["RKESSLER6", "RULS3", "DIENER", "RSBQR", "CDRISC2"]
correlation_matrix = df[scales].corr()
print(correlation_matrix)
plt.figure(figsize=(8,6))
sns.heatmap(correlation_matrix, annot=True, cmap="coolwarm", vmin=-1, vmax=1)
plt.title("Correlation Matrix of Mental Health Scales")
plt.savefig("mental_health_scales_correlation_matrix.png")
```

```
# Variance Influence Factor (VIF) calculation to ensure no multicollinearity among mental health scales
# the highest value was 1.64, which is good, as above 5 indicates high multicollinearity
# which is undesirable for modeling
X_scales = df[scales]
vif_data = pd.DataFrame()
vif_data["feature"] = X_scales.columns
vif_data["VIF"] = [variance_inflation_factor(X_scales.values, i) for i in range(X_scales.shape[1])]
print(vif_data)



# ==============================
# Modeling
# ==============================



# split data into training and testing sets
X_train, X_test, y_train, y_test = train_test_split(
    X, y,
    test_size=0.25,
    stratify=y,
    random_state=1
)

# scale features to make their mean=0 and std=1
scaler = StandardScaler()

X_train_scaled = scaler.fit_transform(X_train)
X_test_scaled  = scaler.transform(X_test)

# ordinal logistic regression model
ordinal_model = OrderedModel(
    y_train,
    X_train_scaled,
    distr="logit"   # logistic link which is standard for ordinal logistic regression
)

ordinal_results = ordinal_model.fit(method="bfgs")

print(ordinal_results.summary())

# Class probabilities
y_pred_prob = ordinal_results.model.predict(
    ordinal_results.params,
    exog=X_test_scaled
)

# Expected ordinal value (better for ordinal outcomes)
classes = np.arange(y_pred_prob.shape[1])
y_pred_expected = np.dot(y_pred_prob, classes)

# evaluate ordinal distance (mean absolute error)
mae = np.mean(np.abs(y_test - y_pred_expected))
print(f"Mean Absolute Error (ordinal): {mae:.3f}")

# evaluate within +/- 1 category
within_one = np.mean(np.abs(y_test - y_pred_expected) <= 1)
print(f"Within +/- 1 category: {within_one:.3f}")

# predicted class = max probability
y_pred = np.argmax(y_pred_prob, axis=1)

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