Peter Gansallo
Dr. Kenton
Due 4-25-25
**FLOAT Method Rationale**

**Formulate**

My research question seeks to explore mental health outcomes among college students, comparing student-athletes and non-student-athletes to identify any differences. Specifically, I aim to investigate how factors such as stigma, social support, loneliness, financial stress, hours of exercise per week, and academic major influence mental health outcomes, such as depression, and anxiety. This question is significant because anxiety and depression are the two most common mental health concerns among college students, and depression and anxiety rates have been on the rise (Zhai et al., 2024). While studies have shown that exercise can positively impact mental health, student-athletes generally engage in more physical activity than non-athletes, but they also face unique stressors related to their academic and athletic commitments (Stansbury, 2023). Stigma has been stated as a barrier that discourages student-athletes from seeking mental health support (Stansbury, 2023), studies measuring stigma and its impact on help-seeking behavior and mental health in this context are scarce. Additionally, research on how a student's major affects their mental health and stigma is also limited. This research seeks to address these gaps by examining how factors which may be big such as exercise, stigma, financial stress and academic major combine to affect the mental health of both student-athletes and non-athletes, and also create a machine learning model to predict students at high risk of depression. The results of this study could inform the development of mental health services tailored to students and student-athletes to contribute to a broader understanding of the factors influencing mental health disparities among college students, and also identify students early at risk of depression.

Previous studies highlight the growing mental health crisis on college campuses, with increasing rates of anxiety, depression, and even suicidal thoughts (Flannery, 2023). Flannery also states that Black and Hispanic students are less likely to seek help for mental health issues. Research state socioeconomic and demographic factors as key influencers of anxiety and depression (Zhai et al., 2024). Stigma has been identified as a limiting factor when comparing mental health outcomes between student-athletes and non-athletes in previous studies. While existing studies have examined mental health outcomes by investigating individual factors, my research aims to build on these findings by exploring the combined effects of multiple factors. In addition, I plan to predict mental health outcomes and classify individuals as low or high risk to identify individuals who may need intervention, which is an area not really explored in this field.

This study will contribute to the existing literature by offering a more comprehensive understanding of how these factors collectively shape mental health on college campuses and by developing a practical tool for identifying students at risk.

**Locate**

For this research, I will be using data from the Health Minds Study, they have been collecting annual survey responses from college students across the U.S. since 2007 from about 500 colleges. This dataset provides a comprehensive set of self-reported mental health data, including multiple likert scale questions, the survey includes PHQ-9 and GAD-7 questions which are clinical standardized questions for screening depression and anxiety. I'll convert these into scores such for anxiety, and depression. Additionally, it includes variables such as stigma, social support networks, loneliness, and demographic information, financial stress, exercise, all of which are important to understanding the mental health of students. The Health Minds Study is well-suited for this research because it's a large dataset containing a lot of qualitative properties, a diverse sample of students from multiple universities, which can provide a broad perspective on the mental health challenges faced by different student populations. This data has many questions in the questionnaire and many variables, it has 1550 variables which makes it a rich source for exploring the research question of how many factors can vary between student-athletes and non-student-athletes.

This dataset is good because it also specifically states if a student is a student athlete or not whereas in other datasets like a UK study dataset in kaggle and a borealis dataset I considered using for analysis didn't have this explicit information. This data set is also good because it has many entries where I can use machine learning to learn from this data. Previous studies have used similar data sources to explore the mental health needs of college students such as NCHA study, the ABCD study, and a Canadian student study by Borealis. This dataset is a valuable resource for expanding upon existing findings by having many variables. By having many variables that directly relate to my research question, and many factors that can influence the outcome of my question I will be able to provide a nuanced assessment of how these factors may contribute to mental health differences among student-athletes and non-athletes.

**Organize**

After choosing the HMS dataset from the various options, I referred to the codebook and questionnaire provided to understand all the variables I was working with. I first focused on cleaning and transforming the data to make it usable for predictive modeling. The dataset contained around 1500 variables, but after reviewing them, I initially selected 143 variables of

interest, narrowing them down further to 40. Given the dataset had approximately 48,000 rows of data, I handled missing values by dropping them and removed columns that didn't significantly affect depression or anxiety scores. For the depression, anxiety, mental wellness, loneliness, support areas, and stigma variables, which were measured using a Likert scale, I combined the related questions into a single score for each category. For the race variable, I simplified it into two categories: white and non-white. I also consolidated the student major information into 8 categories and created a new column called 'major categories.' After organizing the data, I renamed the columns for clarity.

The dataset came encoded, which was beneficial for my analysis, as I wanted to predict students who may be at high risk of anxiety or depression. Following the initial transformations, I further converted the encoded numerical values into their categorical representations to make the data more understandable and good for use in Tableau. This cleaning and transformation process helped me structure the data into a format that could be used in multiple machine learning models and Tableau for understandable visual analysis. My approach was guided by methodologies from machine learning resources and tutorials that informed my feature engineering and data preprocessing steps, and after all the cleaning and organization I was left with a dataset with 7838 rows and 25 columns

**Analyze**

To analyze the dataset, I used Azure Data Explorer to get preliminary looks at the data, and summarized the key variables of interest to see which variables I should keep in the dataset. I focused on depression, anxiety, and mental wellness scores. The summarize function in Kusto was key in generating quick insights and providing statistical measures like averages, minimums, maximums and distributions. These summaries helped identify important trends, like the effect of different demographic and behavioral features on depression scores. By applying statistical tests like t-tests and ANOVA using python in jupyter notebook, I was able to assess the significance of various features, such as financial stress, exercise, and loneliness on mental health outcomes. Additionally, the categorical variables, such as race and student-athlete status, and gender were analyzed to see if there were disparities in mental wellness scores between groups.

For a more visual and interactive exploration, I used Tableau to create multiple visualizations that illustrated the relationships between predictors and depression. The visualizations were key in understanding the data, especially in comparing how depression risk varied across different demographic groups (student-athlete status, financial stress levels, gender). Through this combination of statistical analysis and data visualization, I was able to uncover patterns and identify which variables had the strongest association with depression risk, the biggest 6 factors that increase depression were financial stress, if one was diagnosed with mental illness before, if one

had attended any type of therapy in the last 12 months, loneliness, having eating disorder, and being a student athlete. The biggest 6 factors that decrease risk of depression were exercise hours weekly, GPA, knowing where to receive help, seeking help if needed, being a health or medicine major, and being a business . This provided me with insights that could guide potential interventions or areas for further research to support students' mental health.

**Tell**

To effectively communicate my findings on factors influencing anxiety and depression among college students and student-athletes, I will use a combination of bar charts, box plots, and heatmaps. Bar charts will be used to compare categorical variables, such as gender and student-athlete status,exercise hour groups, with depression risk. Pie charts will help visualize the percentage differences of riskacross different categories. Treemaps will be used to display correlations between key variables, including mental wellness, financial stress, and loneliness, helping to highlight strong relationships between these factors. These visualizations align with my research question by making it easier to identify patterns and disparities in mental health outcomes across different student demographics.

The selected visualizations are the most effective for presenting my findings because they offer clear insights into trends, distributions, and relationships within the data. Bar charts provide a good way to compare group differences, making it easier to see how factors such as student-athlete status or gender impact on depression. Pie charts will help illustrate the percentage differences giving a deeper understanding of how anxiety, depression and other variables vary among students. Treemaps are valuable for showing correlations, helping to identify which factors are most strongly associated with higher mental health risks. Using data-to-viz to find the best visualization choices given my dataset. I believe these choices will provide the best clarity, interpretability, and the ability to highlight key patterns. By using these methods, I ensure that my analysis is both visually compelling and easily understandable to my audience.

**Sources**

Click the link below to see my sources

https://public.tableau.com/views/AnnotatedbibliographyFinal/Dashboard1?:language=en-US&:sid=&:redirect=auth&:display_count=n&:origin=viz_share_link