Methodology

1. Research Overview

This research focuses on inferential analysis, gaining a comprehensive understanding of the association between illicit substance use and mental health issues among young adults that is interpretable to a wider audience. By gaining an interpretable picture of this association, we can clearly communicate results of this research to public health officials and healthcare facilities that are looking to enhance treatment for substance use as well as mental health issues for young adults. This research is also useful for clinicians who are looking to provide timely mental health treatment for those with substance use problems (Han *et al.*, 2022). However, there are many confounding variables that influence the association between illicit substance use and mental health issues among young adults. One example of a confounding variable is impulsivity, which mediates the relationship between substance use and mental health (McHugh *et al.*, 2025). While several confounding variables will be controlled for in our analysis, this study will not establish causality because this is a cross-sectional study that cannot show temporal relationships.

2. Dataset Description

This study uses secondary, public use data (confidential information eliminated) from the 2021-2023 National Survey on Drug Use and Health (NSDUH). The NSDUH data is the leading source of population-based statistical data on behavioral health information like tobacco use, alcohol use, drug use, and mental health. The dataset is cross-sectional and nationally representative of U.S. adolescents and adults. Its unit of observation is the civilian, noninstitutionalized population aged 12 or older in the United States. The NSDUH data was collected with web-based interviews. The sample selected was a state-based, multistage,

stratified area probability sample. The 2021-2023 dataset has about 170,000 rows and 2,600 columns, revealing the importance of narrowing down the data to specific columns and population subsets.

3. Data Cleaning

To collect the data, we narrowed the data down from over 2600 columns to a total of 27 columns, including our independent variables (3), dependent variables (3), covariates (15), and survey design as well as weight variables (See Appendix I). All the columns we selected from the data are imputed (missing values statistically imputed) and recoded variables (derived from one or more edited variables), which almost never contain missing values and are better for analysis. Subsequently, we decoded all the categorical variables from their coded values to descriptive values (e.g. changing 0 to No and 1 to Yes). Then, given that we want overall population estimates for substance use, we replaced values that were 91, 93, 991, or 993 for the imputed substance use variables with 0 because 91, 93, 991, or 993 indicates that they never used the substance in the past 30 days or year. While doing our analysis, to align with our research question, we excluded respondents from the data who are not young adults aged 18-25, making the final sample size 41,873.

4. Analytical Methods

Data was downloaded from the SAMHSA website (Substance Abuse and Mental Health Services Administration, 2025). All statistical analyses will be performed in VS Code software (version 1.104.2, https://code.visualstudio.com/) and R software (version 4.5.1, https://www.r-project.org/). A significance level of $p < 0.05$ will be considered statistically

significant. To account for the complex design of the 2021-2023 NSDUH survey, strata, primary sampling units, and weights will be incorporated anytime we do analysis. For statistical analysis, multivariate, pseudo maximum likelihood (see Appendix II) logistic regression tests, that fit our data types of numerical predictors and binary response, will be applied, which will enable us to see the strength of the substance use and mental health association that has been supported in existing research (Qi et al., 2024). These tests will examine the associations between past-year illicit substance use (marijuana, cocaine, hallucinogens) and past-year suicidal ideation after controlling for sociodemographic characteristics, past-year alcohol use, past-month tobacco use, past-month nicotine vaping, receipt of inpatient treatment, and other mental health indicators (see *Equation*).

$$\log\left(\frac{p(X)}{1-p(X)}\right) = \beta_0 + \beta_1 X_{\text{Marijuana Use}} + \beta_2 X_{\text{Cocaine Use}} + \beta_3 X_{\text{Hallucinogen Use}} + \cdots + \beta_{18} X_{\text{Difficulty Concentrating}}$$

Equation. Adjusted Logistic Regression Model Formulation. ***Note: in creating this formulation, we assume that our predictors are not correlated with each other and there are no interactions.***

Pseudo maximum likelihood logistic regression follows the assumptions of (1) the dependent variable is binary, (2) little or no multicollinearity between predictor variables, (3) model correctly specified (including interactions), (4) linear relationship between predictors and log odds, (5) sufficiently large sample size, (6) complex survey design accounted for, (7) sufficient number of strata as well as PSUs for design-based SEs, and (8) use of design-based variance estimation. To satisfy these assumptions, methods will involve specifying complex survey design, dependent variable recoding, generation of a correlation matrix as well as variance

inflation factors among predictor variables, theory-driven testing for interactions, design-adjusted Wald test for interactions as well as non-linearity, plotting residuals versus fitted values, computing the effective sample size (based on weights), and computing the number of PSUs within each strata. For covariates with more than two categories, dummy variables will be created to ensure they are accounted for in the logistic regression calculation. Two distinct models will be done: an unadjusted model 1 with just main predictors and a fully adjusted model 2 based on all covariates. The logistic regression tests will yield beta coefficients that indicate log odds and 95% confidence intervals. After extracting odds ratios by exponentiating the beta coefficients, if the odds ratio is greater than 1, it can be interpreted as that the odds of the outcome for the exposed group is higher than the odds for the unexposed group. Conversely, if the odds ratio is less than 1, the odds of the outcome for the exposed group is less than the odds for the unexposed group.

References

Han, B., Blanco, C., Einstein, E. B., & Compton, W. M. (2022). Mental health conditions and receipt of mental health care by illicit lysergic acid diethylamide (LSD) use status among young adults in the United States. *Addiction*. https://doi.org/10.1111/add.15789

McHugh, R., McLafferty, M., Brown, N., Ward, C., Walsh, C. P., Bjourson, A. J., McBride, L., Brady, J., O'Neill, S., & Murray, E. K. (2025). The mediating role of impulsivity on suicidal behaviour among higher education students with depression and substance abuse disorders. *Alcohol*, *124*, 89–96. https://doi.org/10.1016/j.alcohol.2025.01.002

Qi, P., Huang, M., & Zhu, H. (2024). Association between alcohol drinking frequency

and depression among adults in the United States: a cross-sectional study. *BMC Psychiatry*, *24*(1). https://doi.org/10.1186/s12888-024-06296-9

Substance Abuse and Mental Health Services Administration. (2025, February 13). *Download NSDUH data files*. SAMHSA. https://www.samhsa.gov/data/data-we-collect/nsduh-national-survey-drug-use-and-health/datafiles/2021-2023

Appendix

I. Data Documentation

Variable Description	Values	Data Type (numeric, categorical, text, date)	% Missingness	Role (Predictor, Covariate, Response
Sex at birth - Imputation Revised	1 = Male, 2 = Female	Categorical (binary)	NA	Covariate
Race/Hispanicity recode	1 = NonHisp White, 2 = NonHisp Black/Afr Am, 3 = NonHisp Native Am/AK Native, 4 = NonHisp Native HI/Other Pac Isl, 5 = NonHisp Asian, 6 = NonHisp more than one race, 7 = Hispanic	Categorical (nominal)	NA	Covariate
Recoded-Education Categories	1 = Less high school, 2 = High school grad, 3 = Some coll/Assoc Dg, 4 = College	Categorical (ordinal)	NA	Covariate

	graduate, 5 = 12 to 17 year olds			
Employment Status 18+ - Imputation Revised	1 = Employed full time, 2 = Employed part time, 3 = Unemployed, 4 = Other (incl. not in labor force), 99 = 12-17 year olds	Categorical (nominal)	NA	Covariate
Private Health Insurance - Imputation Revised	1 = Yes, R does have private health insurance, 2 = No, R does not have private health insurance	Categorical (binary)	NA	Covariate
Recode - Imputation -Revised # Persons in Household	1 = One person in household, 2 = Two people in household, 3 = Three people in household, 4 = Four people in household, 5 = Five people in household, 6 = 6 or more people in household	Categorical (ordinal)	NA	Covariate
RC-Total Family Income Recode	1 = Less than $20,000, 2 = $20,000 - $49,999, 3 = $50,000 - $74,999, 4 = $75,000 or More	Categorical (ordinal)	NA	Covariate
Alcohol Frequency Past Year - Imputation Revised	Range = 1 - 365, 991 = Never Used Alcohol, 993 = Did Not Use Alcohol Past Year	Numerical (discrete)	NA	Covariate

Cig Frequency Past Month - Imputation Revised	Range = 1 - 30, 91 = Never Used Cigarettes, 93 = Did Not Use Cigarettes Past Month	Numerical (discrete)	NA	Covariate
Nicotine Vaping Frequency Past Month - Imputation Revised	Range = 1 - 30, 91 = Never Vaped Nicotine, 93 = Did Not Vape Nicotine Past Month	Numerical (discrete)	NA	Covariate
Binge Alcohol Frequency Past Month - Imputation Revised	Range = 0 - 30, 91 = Never Used Alcohol, 93 = Did Not Use Alcohol Past Month	Numerical (discrete)	NA	Covariate
Recoded-Received Substance Use Treatment As An Impatient - Past Year	0 = No, 1 = Yes	Categorical (binary)	NA	Covariate
How Often Felt Nervous Worst Month in Past Year - Imputation Revised	1 = All of the time, 2 = Most of the time, 3 = Some of the time, 4 = A little of the time, 5 = None of the time, 99 = Legitimate Skip	Categorical (ordinal)	NA	Covariate
How Often Felt Everything Effort Worst Month in Past Year - Imputation Revised	1 = All of the time, 2 = Most of the time, 3 = Some of the time, 4 = A little of the time, 5 = None of the time, 99 = Legitimate Skip	Categorical (ordinal)	NA	Covariate
Difficulty	1 = No	Categorical	NA	Covariate

Concentrating One Month in Past 12 Months - Imputation Revised	difficulty, 2 = Mild difficulty, 3 = Moderate difficulty, 4 = Severe difficulty, 99 = Legitimate Skip	(ordinal)		
Marijuana Frequency Past Year - Imputation Revised	Range = 1 - 365, 991 = Never Used Marijuana, 993 = Did Not Use Marijuana Past Year	Numerical (discrete)	NA	Predictor
Cocaine Frequency Past Year - Imputation Revised	Range = 1 - 365, 991 = Never Used Cocaine, 993 = Did Not Use Cocaine Past Year	Numerical (discrete)	NA	Predictor
Hallucinogen Frequency Past Year - Imputation Revised	Range = 1 - 365, 991 = Never Used Hallucinogens, 993 = Did Not Use Hallucinogens Past Year	Numerical (discrete)	NA	Predictor
Adult Seriously Thought About Killing Self Past Year - Imputation Revised	. = Aged 12-17, 0 = No, 1 = Yes	Categorical (binary)	NA	Response
Adult: Past Year Major Depressive Episode (MDE) - Imputation Revised	. = Aged 12-17, 0 = No, 1 = Yes	Categorical (binary)	NA	Response
Recoded-Received Mental Health Treatment As An Inpatient - Past Year	0 = No, 1 = Yes	Categorical (binary)	NA	Response

Final person-level sample weight	Len: 8 (e.g. 3276.46987)	Numeric (discrete)	NA	NA
Variance stratum	Len: 5 (e.g. 40031)	Numeric (discrete)	NA	NA
Variance primary sampling unit	Len: 2 (e.g. 2)	Numeric (discrete)	NA	NA
Year in which data was collected	Len: 4 (2021-2023)	Numeric (discrete)	NA	NA
Recoded-Only AMI, Only SUD, Both, or Neither - PY-DSM-5-ANY	. = Aged 12-17, 1 = SUD only, no AMI, 2 = AMI only, no SUD, 3 = SUD and AMI, 4 = Neither SUD or AMI	Categorical (nominal)	NA	NA - exploratory purposes
Recoded-Age category	1 = 12-17 Years Old, 2 = 18-25 Years Old, 3 = 26-34 Years Old, 4 = 35 or Older	Categorical (ordinal)	NA	NA - subpop var

II. Pseudo Maximum Likelihood Logistic Regression

a) Context

Pseudo maximum likelihood (PML) is a method used in logistic regression for complex survey data, where standard maximum likelihood estimation (MLE) doesn't work. It extends MLE by incorporating survey weights into the log-likelihood function, adjusting it to account for the sampling design and unequal selection probabilities. This ensures that the resulting parameter estimates represent the target population. PML logistic regression inherently accounts for survey

weights by including them directly in the adjusted log-likelihood, making it suitable for survey data analysis.

b) Equation

$$\ell_w(\boldsymbol{\beta}) \;=\; \sum_{i=1}^{n} w_i \Big[\, y_i \log p_i \;+\; (1 - y_i) \log(1 - p_i) \,\Big].$$

The equation above is the pseudo log-likelihood function for logistic regression. What it intuitively does is figure out the beta parameter (B) that maximizes the log-likelihood function to where the predicted probabilities (Pi) are as close as possible to the outcomes (yi), while accounting for survey weights (wi). These beta parameters, because they are selected based on survey weights, reflect population-level relationships between predictors and the binary outcome. These beta parameters (will be more than one in our case) are then used in interpreting the output of our logistic regression model.