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. Inferential analysis will be done to examine the strength of the association between illicit substance use and mental health as well as to capture potential non-linearities in the association. 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, given that this research is an observational study, we will not be establishing causation.

There are many confounding variables that influence the association between illicit substance use and mental health issues among young adults. For example, let's say a young adult tends to use marijuana frequently and experiences suicidal thoughts on nearly a daily basis. This indicates a positive association between marijuana use and suicidal thoughts. However, it's possible that this adult lives in a poor household and has to work overtime in order to make money for his family. In turn, the adult could use marijuana to cope with the stress of work and also experience suicidal thoughts from his stressful lifestyle. In this case, socioeconomic status is the confounding variable, influencing both marijuana use and suicidal thoughts. Existing research also suggests impulsivity to be a confounder in this association (McHugh *et al.*, 2025). Moreover, while we will account for confounders in my research, we still cannot establish causation as we cannot infer the temporal relationship between illicit substance use and mental

health among young adults due to the data being cross-sectional. Thus, we will only be establishing associations in our research and not causation.

Data Preparation

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, meaning the data was collected at single points in time, and is 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 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. The NSDUH data was collected with web-based interviews. The sample selected was a state-based, multistage, stratified area probability sample.

The dependent variables we selected from this data include if respondent seriously thought about killing self in the past year (0 = No, 1 = Yes), if respondent experienced a major depressive episode in the past year (0 = No, 1 = Yes), and receipt of inpatient mental health treatment in the past year (0 = No, 1 = Yes). Independent variables include marijuana use frequency in the past year (range = 1 - 365, 991 = never used marijuana, 993 = did not use in past year), cocaine use frequency in the past year (range = 1 - 365, 991 = never used cocaine, 993 = did not use in past year), and hallucinogen use frequency in the past year (range = 1 - 365, 991 = never used hallucinogen, 993 = did not use in past year). Covariates include gender, race, education, employment status, health insurance, household size, income, use of other substances

(e.g. alcohol, tobacco, nicotine), receipt of substance use treatment, and other mental health indicators (e.g. how often felt nervous, difficulty concentrating). Finally, we selected variables from the dataset relevant to the survey weights and survey design, including the weights, strata, and primary sampling units (see *Figure 1)*. After excluding respondents who are not young adults aged 18-25, the final sample size is 41,873.

Variable Name	Data Type (numeric, categorical, text, date)	% Missingness	Role (Predictor, Covariate, Response
ANALWT2_C3	Numeric (discrete)	NA	NA - sample weights
VESTR_C	Numeric (discrete)	NA	NA - sample strata
VEREP	Numeric (discrete)	NA	NA - sample PSU
YEAR	Numeric (discrete)	NA	NA - year data collected
IRSEX	Categorical (binary)	NA	Covariate
CATAGE	Categorical (ordinal)	NA	NA - subpop var
NEWRACE2	Categorical (nominal)	NA	Covariate
EDUHIGHCAT	Categorical (ordinal)	NA	Covariate
IRWRKSTAT18	Categorical (nominal)	NA	Covariate
IRPRVHLT	Categorical (binary)	NA	Covariate
IRHHSIZ2	Categorical (ordinal)	NA	Covariate
INCOME	Categorical (ordinal)	NA	Covariate
IRALCFY	Numerical (discrete)	NA	Covariate
IRMJFY	Numerical (discrete)	NA	Predictor
IRCOCFY	Numerical (discrete)	NA	Predictor
IRCIGFM	Numerical (discrete)	NA	Covariate

IRNICVAP30N	Numerical (discrete)	NA	Covariate
IRHALLUCYFQ	Numerical (discrete)	NA	Predictor
IRALCBNG30D	Numerical (discrete)	NA	Covariate
SUTINPPY	Categorical (binary)	NA	Covariate
IRDSTNRV12	Categorical (ordinal)	NA	Covariate
IRDSTEFF12	Categorical (ordinal)	NA	Covariate
IRIMPCONCN	Categorical (ordinal)	NA	Covariate
IRSUICTHNK	Categorical (binary)	NA	Response
IRAMDEYR	Categorical (binary)	NA	Response
MHTINPPY	Categorical (binary)	NA	Response
AMISUD5ANYO	Categorical (nominal)	NA	NA - exploratory purposes

Figure 1. Data Documentation

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. All the columns we selected from the data are imputed (missing values statistically imputed) and recoded (derived from one or more edited variables) 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. Notably, we did NOT filter the data down to young adults aged 18-25 at this

stage (kept all rows of the dataset) because doing so would lead to the survey weights being biased towards that subpopulation.

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, ranging from producing estimates on summary statistics like mean to conducting logistic regression and restricted cubic splines analysis. Demographic, substance use, and mental health patterns among young adults were analyzed using Taylor linearization and tabulation functions from the samplics module in Python, which are functions that produce frequency and mean estimates based on complex survey design. We filtered the data down to our subpopulation when we computed estimates with the samplics module. By gaining a stronger understanding of our sample, we can understand why we see the association that we do between illicit substance use and mental health issues. To ensure that the conditions for taylor linearization and tabulation are met, we calculated the number of PSUs (primary sampling units within each strata), effective sample size ((sum of weights)^2)/(sum of weights^2)), and degrees of freedom (number of strata in our case).

To explore the association between illicit substance use and mental health issues among young adults, we utilized bar plots with error bars (95% confidence intervals) in Python (produced from taylor linearization) and weighted 2 sample t-tests in RStudio to compare illicit

substance use measures (marijuana, cocaine, hallucinogens) across 2 groups: those who experienced X mental health outcome (suicidal thoughts, MDE, receipt of mental health treatment) and those who didn't. If we found a significant difference in average illicit substance use between those with mental health issues and those without, we inferred that there is likely an association between that illicit substance use measure and mental health issues. With this information, we complete the first step to answering our research question: seeing if there is an association between illicit substance use and mental health. To ensure that the conditions for 95% confidence intervals and weighted 2 sample t-tests are met, we calculated the number of PSUs (primary sampling units within each strata), effective sample size ((sum of weights)^2)/(sum of weights^2)), and degrees of freedom (number of strata in our case).

To examine the association between illicit substance use and mental health issues among young adults in depth, logistic regression, restricted cubic splines, and threshold analysis will be utilized. Logistic regression, matching our data types of a numerical predictor and binary response, will enable us to see the strength of the relationship between illicit substance use and mental health (e.g. how much does the chance of major depressive episode change for each one unit increase in marijuana use). Then, restricted cubic splines (RCS) will enable us to see if there are any potential nonlinear associations between illicit substance use and mental health outcomes among young adults, which is used to see the non-linear association between a numerical predictor and categorical response variable. Our research will build on existing research that has shown a non-linear relationship between alcohol use and depression (Qi *et al.*, 2024). The logistic regression tests will yield beta values and 95% confidence intervals.

Before conducting the tests, multicolinearity between covariates will be checked to ensure that we get reliable results, and the complex design of the NSDUH survey will be

accounted for. These tests will involve three distinct models: an unadjusted model 1, a model 2 adjusted for most likely gender, race, and income, and a fully adjusted model 3 based on all covariates. For the first three logistic regression tests, we will select one illicit substance use predictor (most likely past year marijuana use) and one mental health outcome variable (most likely suicidal thoughts). Following these tests, RCS will be employed on the fully adjusted logistic regression model with the end goal of having an interpretable visual derived from the RCS analysis. Finally, threshold effect analysis will be utilized to describe the location of potential inflection points (turning points in the association) by determining the x-values where the derivatives or rates of change are equal to 0. If time permits, we will do logistic regression and RCS as well as threshold analysis for our other predictors and response variables, but only focus on one predictor and one response at a time to ensure the analysis is interpretable.

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