Understanding Student Interest in Substance Use Disorder Counseling: A Mixed-Methods Approach Using Machine Learning and Qualitative Interviews

Erika Hernandez Acton

Bryan Acton

Rory Eckhardt

Linda Reynolds

Giovanni Scaringi

Nadine Mastroleo

Abstract

The demand for Substance Use Disorder (SUD) counselors is rapidly growing, yet significant challenges remain in expanding the workforce to meet this need. Despite substantial investment in recruitment, heterogeneity in student interest hinders pipeline development. To effectively attract and retain future SUD counselors, a deeper understanding of the factors influencing career interest is crucial. This paper presents findings from a two-part, mixed-methods study investigating student interest in SUD counseling. Study 1 employed a survey administered to students at a major state university (N=?) and utilized Random Forest analysis, an advanced machine learning technique, to identify the most salient predictors of interest in pursuing SUD counseling. Study 2 involved qualitative interviews (N=?) exploring students’ perceived barriers, facilitators, and general interest related to SUD and mental health counseling careers. Findings reveal key predictive factors from the quantitative analysis and rich thematic insights from the qualitative data regarding student perspectives. Together, these studies provide critical insights into the complex landscape of student interest in SUD counseling, offering actionable recommendations for targeted workforce development strategies aimed at expanding this vital field.

# Introduction

The field of Substance Use Disorder (SUD) counseling represents a critical component of the behavioral health workforce, addressing a significant public health challenge. National trends indicate a growing demand for qualified SUD counselors, driven by increased awareness, policy changes, and the ongoing opioid crisis (**citation\_needed\_1?**). However, expanding the workforce to meet this demand faces considerable hurdles (**citation\_needed\_2?**). Millions are invested annually in recruitment efforts, yet attracting sufficient numbers of students, particularly those with sustained interest, remains difficult. This difficulty may stem partly from a heterogeneity in student interest levels and a lack of clear understanding regarding the specific factors that draw individuals to, or deter them from, pursuing a career in SUD counseling. Addressing this knowledge gap is essential for developing effective strategies to build and sustain the SUD counseling pipeline. This study employs a mixed-methods approach to identify these critical factors, utilizing both quantitative machine learning techniques (Study 1) and qualitative interviews (Study 2) to provide a comprehensive understanding of student perspectives.

You can cite references like this: (Author 2024). See the paper/references.bib file.

{r setup, include=FALSE} knitr::opts\_chunk$set(echo = FALSE, warning = FALSE, message = FALSE) # Add any R setup code here, like loading libraries # library(tidyverse) # library(readxl)

# Methods

This study employed a mixed-methods approach involving two main components: a quantitative survey analysis using machine learning and qualitative interviews. Ethical approval was obtained from the Institutional Review Board at Binghamton University.

### Data Processing

The raw survey data, collected via Qualtrics and stored in data/survey/survey\_raw.csv, underwent several preprocessing steps to prepare it for analysis using Python libraries, primarily pandas.

First, initial data cleaning was performed by filtering respondents based on survey completion metrics. Only responses with 100% Progress, a Finished status of True, and a CarelessResponderDC value of 0 (indicating the response time exceeded a predefined threshold, likely 120 seconds) were retained for analysis. This filtering resulted in a sample of N=397 participant responses available for subsequent steps.

Next, variables relevant to the analysis were selected and renamed for clarity based on a predefined inclusion list (data/survey/initial\_analysis\_vars\_to\_include.csv). Variables excluded at this stage included survey metadata (e.g., ResponseId, RecordedDate), direct identifiers, free-text responses (e.g., demo\_anythingelse, \_TEXT variables), and specific ranking or qualitative follow-up questions not intended for the primary quantitative analysis. Renaming involved converting original survey variable names (e.g., Duration (in seconds), career\_2) to more programmatic names (e.g., duration\_seconds, sud\_counselor\_interest).

Selected variables were then recoded based on their intended type for analysis, as specified in the inclusion list. Ordinal and Likert scale variables (e.g., sud\_counselor\_interest, sud\_counselor\_familiarity, stress\_\* variables) were converted from their original text descriptions (e.g., “Slightly interested”, “Moderately Stressful”) to numerical values representing the scale levels (e.g., 1, 2, 3, 4, 5), allowing them to be treated as quantitative data where appropriate. For these specific ordinal/Likert variables, responses corresponding to “I prefer not to answer” were recoded as missing data (NaN) to preserve the numeric nature of these scales. In contrast, nominal categorical variables (e.g., gender\_identity, race, current\_year) were retained with their original text values, including “I prefer not to answer” where applicable, as these categories do not have an inherent numerical order. Numerical variables were kept as numbers.

*Handling Sparse Categories:* To address potential model instability and improve interpretability arising from sparse categories in key demographic variables, several categories were grouped prior to imputation and encoding (**citation\_needed\_for\_grouping?**). For race, the categories ‘Black’ (n=12 in training set), ‘Other (please specify):’ (n=11), ‘Middle Eastern’ (n=5), and ‘I prefer not to answer’ (n=2) were combined into a single group labeled ‘Other/Multiple/Unknown Race’. For gender\_identity, the categories ‘I prefer not to answer’ (n=3), ‘Nonbinary’ (n=2), ‘Gender queer’ (n=1), ‘Transgender’ (n=1), and ‘Agender’ (n=1) were combined into a single group labeled ‘Gender Diverse/Unknown’. This grouping step was applied before imputation and one-hot encoding.

Mising values within these predictors were imputed using the median for numeric/ordinal features and the mode for nominal categorical features (after the grouping step described above). Following imputation, the nominal categorical features were converted into numerical format using one-hot encoding, resulting in a final feature set of 99 predictor variables. Validation checks confirmed the absence of missing values and the appropriate numeric data types in the final feature set.

The resulting processed dataset, containing 38 variables for N=397 participants, was saved as data/survey/analysis\_ready\_data.csv. This dataset formed the basis for subsequent descriptive statistics and was the starting point for further processing specific to the machine learning analysis pipeline.

### Study 1: Quantitative Survey and Machine Learning Analysis

**Participants.** Participants were recruited from the undergraduate student population at Binghamton University through the SONA Systems research participation pool. After the initial data processing described above (N=397), participants with missing data on the primary outcome variable (sud\_counselor\_interest recoded as NaN) were excluded, resulting in a final analytical sample of N=391 for the machine learning analysis. The data was then split into training (80%, N=312) and testing (20%, N=79) sets using stratified sampling based on the dichotomized outcome variable. Within the training set (N=312), the distribution for self-reported race included White (n=174), Asian (n=78), Latino or Hispanic (n=30), Black (n=12), Other (n=11), Middle Eastern (n=5), and ‘I prefer not to answer’ (n=2). For gender identity, the training set included Woman (n=228), Man (n=76), ‘I prefer not to answer’ (n=3), Nonbinary (n=2), Gender queer (n=1), Transgender (n=1), and Agender (n=1). The sparsity of several categories within race and gender identity necessitates careful consideration in subsequent modeling steps.

**Measures.** The survey included demographic questions, items assessing interest and familiarity with SUD counseling (original variables career\_1, career\_2), perceived barriers and facilitators (career\_5\_\*, career\_6\_\* - though excluded from ML predictors), knowledge of SUD/MH professionals (career\_3), interest in MH careers (mh\_1), and wellbeing indicators (wellbeing\_\*). See Appendix A for detailed variable descriptions based on the processed data.

**Machine Learning Preprocessing.** Further processing was performed using scikit-learn to prepare the data specifically for the classification model. The primary dependent variable, sud\_counselor\_interest, was dichotomized into a binary variable interest\_dv. Responses indicating any level of interest (“Slightly interested,” “Moderately interested,” or “Definitely interested”) were coded as 1, while responses indicating “Not interested” were coded as 0. The resulting distribution was 62.5% (n=244) in the “Not interested” class (0) and 37.5% (n=147) in the “Any interest” class (1) within the N=391 analytical sample.

Predictor variables included the 34 selected features from the initial processing step (excluding the target variable and filtering columns like progress, careless\_responder, completed). Missing values within these predictors were imputed using the median for numeric/ordinal features and the mode for nominal categorical features. Following imputation, the nominal categorical features were converted into numerical format using one-hot encoding, resulting in a final feature set of 99 predictor variables. Validation checks confirmed the absence of missing values and the appropriate numeric data types in the final feature set.

The fully processed dataset (N=391) was then split into a training set (80%, n=312) and a testing set (20%, n=79). Stratified sampling based on the binary interest\_dv variable was used to ensure that the proportion of interested versus not-interested participants was approximately equal in both the training and testing sets.

**Machine Learning Analysis.** To identify key predictors of student interest in SUD counseling, a Random Forest classification model was employed using the scikit-learn library in Python (**citation\_needed\_1?**). Random Forest, an ensemble method based on multiple decision trees, was chosen for its ability to handle high-dimensional data with mixed predictor types (implicitly, before encoding), capture non-linear relationships between predictors and the outcome, and provide robust estimates of feature importance. The model was trained on the preprocessed training set (n=312) to learn patterns distinguishing between students with any interest versus no interest in SUD counseling.

The performance of the trained model was then evaluated on the independent testing set (n=79). Key performance metrics including overall accuracy, precision, recall, F1-score for each class, and the Area Under the Receiver Operating Characteristic Curve (AUC) will be reported to assess the model’s predictive capability. Additionally, feature importances derived from the trained Random Forest model will be examined to identify the most influential factors associated with expressing interest in pursuing an SUD counseling career.

While several models showed reasonable performance, the **Logistic Regression model utilizing L1-based feature selection** emerged as the preferred choice. L1 regularization (also known as Lasso) was employed because it performs automatic feature selection by adding a penalty proportional to the absolute value of the coefficients. This process shrinks the coefficients of less important features towards zero, potentially setting some exactly to zero, thus removing them from the model. This approach leads to a more parsimonious model (one with fewer predictors) which is often easier to interpret, a key goal for understanding the drivers of student interest. Applying L1 regularization during the tuning process identified a subset of 18 potentially relevant features (out of the initial 99). A final Logistic Regression model was then trained and tuned using only these 18 selected features. This model achieved the highest test set ROC AUC score (0.821), offered a good balance between precision and recall, and provided greater simplicity compared to more complex models like Random Forest or XGBoost.

(Note: Detailed performance metrics, confusion matrices, and coefficient analyses for all explored models, including Random Forest, XGBoost, SVM, the final binary Logistic Regression, and an exploratory Ordinal Logistic Regression, are available in the project’s results/ directory and can be provided as supplementary material.)

Performance of the final feature-selected Logistic Regression model on the test set is detailed below. [Figure 1](#fig-confusion-matrix) shows the confusion matrix, illustrating the counts of true positives, true negatives, false positives, and false negatives. [Figure 2](#fig-roc-curve) presents the ROC curve, visually representing the trade-off between the true positive rate and false positive rate, with an overall AUC of 0.821.

|  |
| --- |
| Figure 1: Confusion Matrix for Feature-Selected Logistic Regression on Test Set. |

|  |
| --- |
| Figure 2: ROC Curve for Feature-Selected Logistic Regression on Test Set (AUC = 0.821). |

To understand which factors were most influential in the final model, **?@tbl-coefficients** presents the top 15 features ranked by the absolute magnitude of their coefficients. The odds ratio indicates the multiplicative change in the odds of being interested in an SUD counseling career for a one-unit change in the predictor, holding other predictors constant. Odds ratios greater than 1 suggest a positive association, while those less than 1 suggest a negative association.

```{python}  
#| label: tbl-coefficients  
#| tbl-cap: "Top 15 Features by Absolute Coefficient Magnitude in Final Logistic Regression Model."  
#| echo: false  
import pandas as pd  
  
# Use relative path from the QMD file in the root directory  
coeffs\_path = 'results/study1\_logistic\_fs/logistic\_fs\_coefficients.csv'  
try:  
 coeffs\_df = pd.read\_csv(coeffs\_path)  
   
 # Select top 15 features based on absolute coefficient  
 top\_coeffs = coeffs\_df.iloc[coeffs\_df['coefficient'].abs().argsort()[::-1]]  
 top\_15\_coeffs = top\_coeffs.head(15)  
   
 # Select and rename columns for the table  
 table\_df = top\_15\_coeffs[['feature', 'coefficient', 'odds\_ratio']].copy()  
 table\_df.rename(columns={  
 'feature': 'Feature',  
 'coefficient': 'Coefficient',  
 'odds\_ratio': 'Odds Ratio'  
 }, inplace=True)  
   
 # Round numeric values for presentation  
 table\_df['Coefficient'] = table\_df['Coefficient'].round(3)  
 table\_df['Odds Ratio'] = table\_df['Odds Ratio'].round(3)  
   
 # Print as markdown table  
 print(table\_df.to\_markdown(index=False))  
   
except FileNotFoundError:  
 print(f"Error: Coefficient file not found at {coeffs\_path}")  
except Exception as e:  
 print(f"Error processing coefficients: {e}")  
```

```{python}  
#| label: fig-feature-importance  
#| fig-cap: "Coefficients of the 18 Features Selected by L1 Regularization in the Final Logistic Regression Model. Positive coefficients indicate increased odds of interest in SUD counseling, while negative coefficients indicate decreased odds."  
#| echo: false  
import pandas as pd  
import matplotlib.pyplot as plt  
import numpy as np  
  
# Use relative path from the QMD file in the root directory  
coeffs\_path = 'results/study1\_logistic\_fs/logistic\_fs\_coefficients.csv'  
try:  
 coeffs\_df = pd.read\_csv(coeffs\_path) # Contains all 18 selected features  
   
 # Sort by coefficient value for plotting  
 coeffs\_df\_sorted = coeffs\_df.sort\_values('coefficient', ascending=False)  
   
 # Create the plot  
 fig, ax = plt.subplots(figsize=(8, 6)) # Adjust figsize as needed  
   
 colors = ['#0072B2' if c > 0 else '#D55E00' for c in coeffs\_df\_sorted['coefficient']] # Example colors for pos/neg  
   
 ax.barh(coeffs\_df\_sorted['feature'], coeffs\_df\_sorted['coefficient'], color=colors)  
   
 # Add vertical line at zero  
 ax.axvline(0, color='grey', linewidth=0.8, linestyle='--')  
   
 # Improve labeling and aesthetics  
 ax.set\_xlabel('Coefficient Value')  
 ax.set\_ylabel('Feature')  
 ax.set\_title('Feature Importance and Direction')  
 ax.invert\_yaxis() # Display features with highest positive coeffs at top  
 plt.tight\_layout() # Adjust layout to prevent labels overlapping  
 plt.show() # Display the plot  
  
except FileNotFoundError:  
 print(f"Error: Coefficient file not found at {coeffs\_path}")  
 # Optionally, create a placeholder plot or message  
 fig, ax = plt.subplots()  
 ax.text(0.5, 0.5, 'Coefficient data not found,\ncannot generate plot.',   
 ha='center', va='center', fontsize=12, color='red')  
 ax.set\_xticks([])  
 ax.set\_yticks([])  
 plt.show()  
except Exception as e:  
 print(f"Error generating feature importance plot: {e}")  
 # Optionally, create a placeholder plot or message  
 fig, ax = plt.subplots()  
 ax.text(0.5, 0.5, f'Error generating plot:\n{e}',   
 ha='center', va='center', fontsize=10, color='red')  
 ax.set\_xticks([])  
 ax.set\_yticks([])  
 plt.show()  
  
```

The factors most strongly influencing student interest are visualized in **?@fig-feature-importance**, with specific coefficients and odds ratios for the top 15 detailed in **?@tbl-coefficients**. Notably, higher familiarity with the SUD counselor profession (sud\_counselor\_familiarity, OR > 1) and identifying as Hispanic/Latino (race\_Hispanic / Latino) were associated with significantly increased odds of expressing interest. Conversely, factors such as being in the senior year (current\_year\_Senior) or identifying with certain other racial categories (e.g., race\_White, race\_Asian) were associated with decreased odds compared to the reference groups used in the encoding. Wellbeing factors, such as stress related to housing stability (stress\_housing\_stability), also appeared among the influential predictors, though their interpretation requires care given the nature of Likert scales. Further investigation, potentially integrating qualitative findings, is needed to fully understand the complex interplay of these demographic, experiential, and wellbeing factors in shaping career interest.

The specific hyperparameters used for this final model were C=1000, class\_weight='balanced', penalty='l1', and solver='liblinear'. The full classification report and details for all models explored are available in the project’s results/ directory.

### Study 2: Qualitative Interviews

**Participants.** A subset of survey respondents who indicated willingness to participate in a follow-up interview were contacted. [Add details on final interview sample size, recruitment criteria, demographics].

**Procedure.** Semi-structured interviews were conducted via Zoom, lasting approximately 30-45 minutes. Interviews explored students’ understanding of SUD/MH counseling, perceived barriers and facilitators to entering these fields, and factors influencing their career interests.

**Analysis.** Interview transcripts were analyzed using thematic analysis (**citation\_needed\_2?**) to identify recurring patterns and themes related to student perspectives on SUD and MH counseling careers.

# Results

To predict student interest in pursuing a career related to substance use disorder (SUD) counseling (interest\_dv), several machine learning models were developed and evaluated using data collected from the survey. The primary goal was to identify a model with strong predictive performance, measured primarily by the Area Under the Receiver Operating Characteristic Curve (ROC AUC), while also considering other metrics like accuracy, precision, recall, and F1-score, particularly for the positive class (students expressing interest).

Models explored included Random Forest (with and without SMOTE oversampling), Logistic Regression (with L1 and L2 regularization), Support Vector Machine (SVM), and XGBoost. Hyperparameter tuning was performed using cross-validation, and models were evaluated on a held-out test set.

# Discussion

Discuss the implications of your results, integrating findings from both studies. Consider how the qualitative findings might elaborate on or provide context for the quantitative predictors.

# References

Author, A. N. 2024. “Example Citation for Demonstration.” Publisher.

# Appendix

## Variable Descriptions

| Variable | Description | Type |
| --- | --- | --- |
| Progress | The variable ‘Progress’ measures the percentage of completion or advancement in a given task or project, represented as a numerical value from 0 to 100. | Numeric |
| Duration (in seconds) | This variable measures the duration of an event or activity in seconds, capturing how long it takes to complete a specific task. | Numeric |
| CarelessResponderDC | This variable indicates whether a respondent completed the survey in less than the threshold duration of 120 seconds, with a value of 1 signifying a potentially careless response. | Categorical |
| Finished | The variable ‘Finished’ indicates whether a task or activity has been completed, with a binary response of either completed (True) or not completed (False). | Categorical |
| RecordedDate | The RecordedDate variable captures the date and time when a particular event or response was logged, providing a timestamp for data collection. | Other |
| ResponseId | Response ID is a unique identifier assigned to each survey response, allowing for tracking and analysis of individual submissions. | Text |
| Q46 | This variable measures the consent status of participants regarding their age and willingness to participate in the research study. | Categorical |
| demo\_age | This variable measures the age of respondents in years, capturing a range of age groups and preferences regarding age disclosure. | Categorical |
| demo\_gender | This variable measures the gender identity of respondents, capturing their self-identification in terms of gender. | Categorical |
| demo\_sex | This variable measures the sex assigned to an individual at birth, reflecting their biological classification. | Categorical |
| demo\_country | This variable captures the country of birth of the survey respondent, providing insights into demographic backgrounds. | Categorical |
| demo\_race | This variable captures the racial identity of respondents as part of demographic data collection. | Categorical |
| demo\_race\_7\_TEXT | This variable captures the self-reported race of respondents who selected ‘Other’ in a survey, allowing for open-ended text responses to specify their racial identity. | Text |
| demo\_served | This variable measures whether an individual has ever served on active duty in the U.S. Armed Forces, indicating their military service status. | Categorical |
| demo\_disability | This variable measures whether an individual has a formally diagnosed disability as recognized by a medical professional. | Categorical |
| demo\_schoolyear | This variable measures the current year in school of the respondents, indicating their level of progression in their educational journey. | Categorical |
| demo\_parenteducation | This variable measures the highest level of education completed by the respondent’s parents or guardians, reflecting their educational background. | Categorical |
| demo\_employment | This variable measures the current employment situation of respondents, capturing whether they are employed, unemployed, or in school with varying work hours. | Categorical |
| demo\_housing | The variable measures the current living situation of respondents while attending Binghamton University, indicating whether they reside on-campus, off-campus, or with family. | Categorical |
| demo\_livewith | This variable measures the total number of friends or family members living with the respondent at their current residence while attending school. | Numeric |
| demo\_safety | This variable measures the respondents’ perception of their physical safety in their neighborhood while attending school. | Categorical |
| demo\_permanenthome | This variable measures the type of permanent residence of the respondent when not attending school, indicating their living situation. | Categorical |
| demo\_permanenthome\_5\_TEXT | This variable captures the description of the respondent’s permanent residence when not attending school, allowing for open-ended responses. | Text |
| demo\_geography | This variable measures the type of geographic area in which the respondent grew up, categorizing their upbringing into distinct environments. | Categorical |
| demo\_safeathome | This variable measures the respondent’s perception of physical safety in their childhood neighborhood, reflecting their feelings of security during that time. | Categorical |
| demo\_caregiver | This variable measures whether the respondent serves as a caregiver for individuals aged 18 or older. | Categorical |
| demo\_familyincome | This variable measures the annual household income of respondents, capturing the combined income of all individuals living in their home or permanent residence. | Categorical |
| demo\_personalincome | This variable measures the respondent’s personal annual income, capturing a range of income levels as well as options for non-disclosure. | Categorical |
| demo\_religion | This variable measures the respondent’s religious affiliation or beliefs, capturing their identification with specific religious branches or lack thereof. | Categorical |
| demo\_addiction | This variable measures whether an individual has ever been diagnosed or treated for a substance use or addiction concern, providing insight into their personal history with addiction. | Categorical |
| demo\_familyaddiction | This variable measures whether a respondent has a close friend or family member who has been diagnosed or treated for a substance use or addiction concern, indicating the prevalence of addiction issues within personal networks. | Categorical |
| demo\_mentalhealth | This variable measures whether an individual has ever been diagnosed or treated for a mental health concern, indicating their mental health history. | Categorical |
| demo\_people | This variable measures the frequency with which individuals engage in social interactions with people they care about, reflecting their social connectivity and support network. | Ordinal/Likert |
| demo\_anythingelse | This variable captures additional information about the respondent’s background that may not be covered by other survey questions, allowing for open-ended responses. | Text |
| career\_1 | This variable measures the respondent’s level of familiarity with the substance use disorder counselor profession, ranging from no familiarity to a high degree of familiarity. | Ordinal/Likert |
| career\_2 | This variable measures the respondent’s level of interest in pursuing a career as a substance use disorder counselor. | Ordinal/Likert |
| career\_4\_1 | This variable measures the ranked importance of various reasons for interest in becoming a substance use disorder counselor, focusing on personal motivations and values. | Ordinal/Likert |
| career\_4\_2 | This variable measures the ranked importance of various reasons for interest in becoming a substance use disorder counselor, specifically focusing on the second reason selected by respondents. | Ordinal/Likert |
| career\_4\_3 | This variable measures the ranked importance of various reasons for interest in becoming a substance use disorder counselor, focusing on personal motivations and values. | Ordinal/Likert |
| career\_6\_1 | This variable measures the ranked importance of factors that influence the interest in pursuing a career as a substance use disorder counselor, with respondents identifying their top three priorities. | Ordinal/Likert |
| career\_6\_2 | This variable measures the ranked importance of factors that could make a career as a substance use disorder counselor more appealing or feasible for respondents. | Text |
| career\_6\_3 | This variable captures the ranked preferences of respondents regarding factors that would enhance the appeal or feasibility of a career as a substance use disorder counselor. | Text |
| career\_5\_1 | This variable captures the top three reasons respondents are not interested in pursuing a career as a substance use disorder counselor, ranked by importance. | Text |
| career\_5\_2 | This variable captures the top three reasons respondents are not interested in pursuing a career as a substance use disorder counselor, reflecting personal motivations and barriers. | Text |
| career\_5\_3 | This variable captures the top three reasons respondents are not interested in pursuing a career as a substance use disorder counselor, reflecting personal attitudes and experiences related to the field. | Text |
| career\_3 | This variable measures the respondent’s awareness of individuals who have worked or are currently working as substance use disorder counselors. | Categorical |
| career\_3fu | This variable measures the relationships of respondents to individuals who have worked or are currently working as substance use disorder counselors, capturing personal connections such as family members or friends. | Text |
| mh\_1 | This variable measures the respondent’s interest in pursuing a career in mental health counseling or related fields. | Categorical |
| mh\_4 | This variable measures the areas of specialization that respondents are interested in pursuing if they were to become a mental health counselor, allowing for multiple selections. | Categorical |
| mh\_4\_12\_TEXT | This variable captures the areas of specialization that respondents are interested in pursuing if they were to become mental health counselors, specifically focusing on ‘other’ specialties that are not listed in predefined options. | Text |
| mh\_1.5 | This variable captures the various classes and training experiences that respondents believe are necessary for individuals to become mental health counselors or therapists. | Text |
| Q44 | This variable captures the names and titles of degree programs that respondents associate with becoming a mental health counselor, reflecting their perceptions of relevant educational pathways. | Text |
| Q45 | This variable measures respondents’ perceptions of the duration required to become a mental health counselor, reflecting their understanding of the training and education involved. | Categorical |
| Q46.1 | This variable measures the perceived cost of education or training required to become a mental health counselor, capturing a range of monetary values and uncertainty. | Categorical |
| Q47 | This variable measures the preferred location for attending a training program in mental health counseling among respondents. | Categorical |
| mh\_3fu | This variable captures the reasons respondents prefer a specific location, reflecting their personal experiences and motivations related to that location. | Text |
| wellbeing\_1 | This variable measures the level of concern or stress individuals experience regarding education-related expenses, capturing their subjective feelings about this financial issue. | Ordinal/Likert |
| wellbeing\_2 | This variable measures the level of concern or stress individuals experience regarding the cost of housing. | Ordinal/Likert |
| wellbeing\_3 | This variable measures the level of concern or stress related to the stability of housing, as reported by respondents. | Ordinal/Likert |
| wellbeing\_4 | This variable measures the level of concern or stress individuals experience regarding the cost of groceries. | Ordinal/Likert |
| wellbeing\_5 | The variable measures the level of concern or stress individuals experience regarding the cost of transportation. | Ordinal/Likert |
| wellbeing\_6 | This variable measures the level of concern or stress individuals experience regarding the reliability of transportation. | Ordinal/Likert |
| wellbeing\_7 | This variable measures the level of concern or stress individuals experience regarding access to high-speed internet. | Ordinal/Likert |
| wellbeing\_8 | This variable measures the level of concern or stress individuals experience regarding the cost of childcare. | Ordinal/Likert |
| wellbeing\_9 | This variable measures the level of concern or stress individuals experience regarding access to childcare. | Ordinal/Likert |
| wellbeing\_10 | The variable measures the level of concern or stress individuals feel regarding the stability and safety of their relationships with people they currently live with. | Ordinal/Likert |

## Example Analysis Code

Example code for loading data and generating outputs from Study 1.

Loading processed data and displaying a table:

```{python}  
# import pandas as pd  
# df = pd.read\_csv("../data/processed/cleaned\_data.csv")  
# print(df.head().to\_markdown(index=False)) # Example using to\_markdown for Quarto tables  
```

Including a figure:

```{python}  
# import matplotlib.pyplot as plt  
# import numpy as np  
# plt.figure()  
# plt.plot(np.random.rand(10))  
# plt.title("Example Figure")  
# plt.show() # Quarto captures the plot  
```