***Data Science Lab  
Data Processing Report***

*Lour Hakim – 323038091  
Shahar Rashdi – 314948415*

**Mental Health in U.S. High School Students:   
Exploring Risk Factors and Economic Context through YRBS and GDP Data**

***Data Schema, Sources, and Curation***

*Our project integrates three distinct datasets to investigate mental health risk factors among U.S. high school students. The primary dataset is the 2023 Youth Risk Behavior Survey (YRBS), an official CDC survey that includes over 100 student-level variables related to emotional well-being, behaviors, demographic and school life. To incorporate economic context, we merged two additional datasets: (1) GDP by state from the U.S. Bureau of Economic Analysis, and (2) unemployment rate and mean household income from the U.S. Census Bureau’s American Community Survey. The latter was constructed from eight state-level CSV files.*

*Entities are linked through the sitename variable, which maps each student to their district and corresponding state. This enabled us to enrich student-level information with macro-level socioeconomic indicators. During schema curation, we removed non-informative columns (e.g., sitecode, columns with only one unique value) and discarded metadata unrelated to prediction. The resulting structure allowed us to analyze cross-level predictors of adolescent mental health.*

***Statistical Population and Variables***

*Each data point in our dataset represents a single U.S. high school student who completed the 2023 Youth Risk Behavior Survey (YRBS). This nationwide survey employs a stratified, cluster-based sampling design: schools and classrooms are randomly selected, and students anonymously complete the questionnaire during class. The responses are statistically weighted to reflect the broader high school population across the United States. We treat each student record as an independent and identically distributed (i.i.d.) observation, enabling generalization to the national level.*

*All variables in the dataset are treated as random variables. This includes individual-level features (e.g., age, sex, bullying, sleep) as well as state-level socioeconomic indicators (e.g., GDP, unemployment rate, household income), which are mapped to each student via the sitename field. These variables jointly reflect behavioral, demographic, and contextual influences on youth mental health.*

*Our analysis targets three mental health outcomes:*

* ***Q26****: “Felt sad or hopeless for two or more weeks” – binary*
* ***Q27****: “Seriously considered suicide” – binary*
* ***Q84****: “Days of poor mental health in past 30 days” – ordinal with five levels: Never, Rarely, Sometimes, Most of the time, Always*

*Variables in the dataset fall into three types: categorical (e.g., sex, race), ordinal (e.g., Q84, breakfast frequency), and numeric (e.g., age, BMI, GDP). We do not assume normality or linearity in the data. Instead, we allow for complex and potentially non-linear interactions between variables. This flexible view supports the modeling goal of identifying risk factors that influence adolescent well-being across diverse personal and socioeconomic settings.*

***Exploratory Data Analysis***

* *We began by examining the distribution of key individual variables.*
* *Mental health outcomes (Q26, Q27, Q84) showed that ~38% of students felt sad or hopeless, ~17% had suicidal thoughts, and emotional distress (Q84) was widespread.*
* *Demographics such as sex, race, grade, and age showed balanced representation with Hispanic/Latino being the largest group.*
* *Behavioral variables such as bullying (Q24) and breakfast (Q75) displayed meaningful spread — ~17% reported being bullied, and breakfast consumption varied widely.*
* *Health variables (BMI) showed a right-skewed distribution with a long tail, indicating the presence of high-value outliers.*
* *Socioeconomic indicators (GDP, unemployment, income) varied substantially across states, suggesting useful context-level variation.*
* *We investigated associations between variables using Chi² tests and Cramér’s V.*
* *Bullying and sex had the strongest relationships with sadness (Q26), with Cramér’s V values of 0.141 and 0.161 respectively.*
* *BMI outliers also showed a significant association with suicidal thoughts.*
* *We identified moderately redundant variables such as physical violence and neighborhood safety but retained them due to distinct contexts.*
* *Before applying complex models, we evaluated simple baselines:*
* *A constant classifier predicting “No sadness” achieved 61.5% accuracy.*
* *A univariate logistic regression using only Q24 (bullying) improved performance to 65.5%, indicating strong predictive value even for a single variable.*
* *Visualizations across the presentation (e.g., distributions, cross-tabs, grouped bar charts) clearly illustrated the patterns:*
* *Students bullied at school showed much higher rates of reported sadness.*
* *Suicidal thoughts were more common among students flagged as BMI outliers.*
* *Emotional distress varied by state.*
* *We used formal statistical tests to validate findings:*
* *Chi² test comparing BMI outliers and suicidal thoughts yielded χ² = 31.33, p < 0.001, indicating a significant association.*
* *Cramér’s V values were computed to assess variable informativeness.*
* *These results guided variable selection for modeling.*
* *Exploratory data analysis and sparse logistic regression played complementary roles in shaping our project design. The EDA phase revealed significant relationships between mental health outcomes and predictors such as bullying, BMI, sleep, breakfast habits, and perceived unfairness. These findings led to the creation of derived features (e.g., bmi\_outlier) and the prioritization of behaviorally and demographically relevant variables for modeling.  
  To extend this, we trained a sparse logistic regression model using L1 regularization, which highlighted the top non-zero coefficients among over a hundred predictors. Key contributors included state-level GDP, unemployment rate, bullying, income, sleep, and sexual and gender identity. This approach confirmed many EDA insights while surfacing additional influential features (e.g., non-prescribed medication use, feeling close to others at school) that had not been deeply analyzed in univariate steps.  
  At this stage, we have not yet removed redundant or overlapping variables. However, the combination of EDA and model-driven feature selection has provided a clear path forward. In the next phase, we will refine our variable set by filtering highly correlated or collinear predictors and consolidating overlapping constructs. These steps will help streamline the model, improve generalizability, and retain only the most interpretable and predictive features.*

***Cleaning***

*We applied model-agnostic cleaning and transformation steps to ensure data quality for analysis. Rows with more than 70% missing values were removed. Numeric variables (e.g., BMI) were imputed using the mean, and categorical variables were filled with -1 to indicate non-response. Outliers in BMI were flagged using the IQR method (thresholds <7.6 or >37.2) and labeled using a binary variable (bmi\_outlier). These were retained, as they represent valid but extreme values and do not appear to be irrelevant or misleading. We also identified potentially dishonest responses from students, which may introduce additional noise, and treated these with caution.*