MINISTRY OF EDUCATION AND TRAINING

NATIONAL ECONOMICS UNIVERSITY

THE CROSSROADS OF GROWTH – WHEN THE DREAM OF EDUCATION BECOMES A PSYCHOLOGICAL BURDEN IN INDIA

Course: Data Preparation & Visualization

Lecturer: Dr. Nguyen Tuan Long

	Full Name	Student ID
1	Nguyễn Thị Vân Anh	11230603
2	Nguyễn Mạnh Cường	11230521
3	Trần Thu Hiền	11230534
4	Lý Thành Long	11230561
5	Nguyễn Thanh Mơ	11230571

Hanoi, December 8, 2025

TABLE OF CONTENT

I. OVERVIEW OF DATA STORYTELLING APPROACH

1. Objectives of Storytelling in the Project

Data storytelling is an essential framework in modern data science. It functions not only as a communication tool but also as a structured analytical lens that helps examine how raw data evolves into meaningful insights. In this project - *"The Crossroads of Growth: When the Dream of Education Becomes a Psychological Burden in India"* - storytelling is used to illuminate how an unrefined dataset on mental-health factors can be transformed into a coherent, interpretable, and analytically sound foundation for predictive modeling.

The core aim of the storytelling framework is to construct a compelling, evidence-based narrative that highlights the substantive difference between raw data and data that has undergone systematic preprocessing. Through this approach, the project emphasizes a central principle: predictive success in mental-health analytics depends fundamentally on the quality of data preparation. Although raw survey responses capture respondents' lived experiences, they do not inherently support accurate pattern detection. Only through disciplined preprocessing can the dataset achieve the structural clarity required to reveal meaningful relationships among behavioral, psychosocial, and environmental determinants of depression.

This narrative-driven approach is designed to achieve several outcomes:

First, it facilitates deeper and more interpretable insights into the determinants of depression, helping researchers separate genuine signals from noise in a domain marked by subjectivity and multi-layered complexity.

Second, it strengthens model performance by providing machine-learning algorithms with cleaner, more consistent, and statistically aligned inputs.

Third, storytelling supports the construction of visualizations that are both accessible and faithful to the psychological constructs represented in the data, reducing the ambiguity commonly associated with unprocessed survey information.

Taken together, these outcomes demonstrate how higher-quality data directly improves interpretability, predictive accuracy, and the overall credibility of analytical findings for both student and working-professional populations.

2. Importance of Data Preparation

The importance of data preparation becomes particularly evident in the context of our depression-prediction project, where self-reported variables often exhibit inconsistency, subjectivity, and considerable heterogeneity. Psychological survey data naturally contains noise arising from differences in self-awareness, varying interpretations of scaled questions, and fluctuating emotional states at the time of response. Additional issues - including logical inconsistencies and missing values - were also present in the dataset.

For these reasons, data preparation is indispensable in mitigating errors and ensuring analyses capture genuine behavioral and emotional patterns. Cleaning procedures such as imputing missing values,

harmonizing categorical formats, handling outliers, and resolving contradictory entries provide the necessary foundation for stable and reproducible model training. Transformation steps such as scaling and encoding further enhance the dataset's analytical integrity by aligning numerical features to comparable ranges, improving gradient-based model convergence, and ensuring categorical attributes are expressed in statistically meaningful forms.

The raw depression dataset used in this project illustrates these challenges clearly. It contained missing responses in critical variables such as Suicidal Thoughts and Financial Stress - issues that, if ignored, could bias model estimates or reduce sensitivity for high-risk subgroups. Noise appeared through inconsistent use of rating scales, reporting errors, and abrupt numerical jumps in stress-related measures. Outliers - particularly extreme stress levels or irregular sleep durations - threatened to distort distributional properties and amplify model error. Additionally, weakly related or irrelevant features risked introducing unnecessary variance, while large scale differences across variables hindered the model's ability to learn meaningful depression-related patterns.

These limitations underscore that data preparation is not an optional step but a foundational component of the analytical process - one that ensures the resulting models are both accurate and trustworthy.

II. STORY BUILDING: FROM RAW DATA TO A RELIABLE MENTAL HEALTH NARRATIVE

1. WHO - Who are the subjects, and why does this story matter?

1.1. Research Subjects

The study focuses on two primary groups who differ in characteristics and psychological contexts. The first group is Students, consisting of individuals aged 18 to over 30 who are influenced by academic pressure, expectations for clear career direction, and financial dependence on family or external support. The second group is Workers, ranging from 18 to 60 years old, who face the persistent challenges of the labor environment, such as job pressure, family responsibilities, and the need to maintain long-term financial stability. Dividing the population into these two groups allows the research to assess differences in depression-related factors across distinct psychosocial developmental pathways.

1.2. Beneficiaries of the Research (Audience)

The research findings aim to provide value to multiple relevant stakeholders. First are adolescents and young adults who are preparing to enter adulthood and need to understand potential psychological risks in order to adopt appropriate preventive strategies. In addition, the analytical results serve as an important reference for educational institutions, employers, mental health support teams, and psychological researchers who directly participate in designing or implementing mental health care programs. Moreover, organizations and institutions responsible for developing mental health policies or intervention programs can use the study's insights to build tailored solutions for each target group.

1.3. Problems They Are Facing

Both Students and Workers are experiencing increasing rates of depression, while the specific contributing factors are difficult to determine due to substantial differences in living environments, personal pressures, and social conditions. This issue becomes more serious when raw mental health data are often incomplete, containing noise, missing values, or distortions that undermine analytic reliability. As a result, the insights derived may be biased, leading to inappropriate support policies or ineffective psychological interventions. Practical and reliable solutions cannot be developed without a clear understanding of the true nature of the data and the actual influence of each factor.

1.4. The Core of the Storytelling Approach

In the context of increasingly complex mental health data, Data Preparation plays a crucial role in ensuring the accuracy and consistency of analytic outcomes. Only when data are cleaned, standardized, and systematically processed can we distinguish between the core determinants of depression and mere noise. Storytelling through data is meaningful only when the data foundation is sufficiently robust; therefore, to accurately uncover and interpret the factors influencing depression among Students and Workers, Data Preparation must be regarded as an indispensable step that establishes the groundwork for the entire research process.

2. WHAT - What problem are we addressing, and what did we find?

2.1. Research Problem

The study focuses on comparing and identifying the causes underlying differences in depression rates between two groups of adults who have chosen two major developmental pathways: pursuing higher education (Students) and participating in the workforce (Workers). These groups differ significantly in living environments, personal pressures, and social expectations; therefore, the factors contributing to depression are likely to vary according to the characteristics of each group. The analysis aims to clarify which factors exert the strongest influence on each group and why depression rates may diverge between these two developmental trajectories.

Preliminary results show that the raw data do not accurately reflect the actual correlations among variables. The presence of noisy, missing, or irregularly distributed data distorts initial conclusions, highlighting the importance of conducting systematic Data Preparation before drawing any insights related to mental health.

2.2. Core Research Questions

Based on the research objectives, several key questions are posed to guide the analytical process and determine the nature of differences in depression rates between the two groups. First, it is necessary to clarify which factors influence depression rates among individuals choosing different pathways for their future development (continuing higher education vs. entering the workforce), and to identify the similarities and differences between these contributing factors.

Next, the study examines the specific variables that increase the risk of depression among students, including those related to academic demands, financial conditions, social relationships, and behavioral characteristics. At the same time, the analysis evaluates the factors influencing workers, particularly those arising from the work environment, family responsibilities, and long-term financial pressures.

Finally, a methodological question is raised: What new or different insights does Data Preparation produce compared with analyzing raw data directly? Addressing this question clarifies the role of data processing in removing noise, correcting distortions, and ensuring that the conclusions reflect the true nature of the issue under investigation.

2.3. Key Findings After Data Preparation

Following the processes of cleaning, standardization, and data restructuring, the analytical models identified the factors influencing depression more clearly for both Students and Workers. For Students, the results indicate that psychological distress stemming from unresolved emotional burdens, combined with academic-specific pressures, constitutes the strongest drivers of depression risk. Additionally, factors such as social comparison and financial dependence were found to exacerbate stress levels, creating a detrimental cycle for mental well-being.

Among workers, the models show that age-related pressure-particularly comparisons with peers of the same age who report higher job satisfaction-combined with lower levels of job satisfaction, significantly affects depression levels. Furthermore, family responsibilities and long-term financial burdens were identified as important and persistent factors exerting strong psychological impact on

working individuals.

The distinction between the two groups becomes clearer when examining the nature of these contributing factors. Students are influenced primarily by endogenous elements, including self-perception, personal expectations, and peer pressure. In contrast, workers are strongly affected by exogenous factors such as job-related stress, financial obligations, and social roles. This contrast reflects two different pathways of adulthood, each shaped by distinct mental health challenges.

2.4. Unexpected Insights Revealed Only After Data Preparation

Notably, many correlations initially considered "significant" in the raw dataset disappeared completely after technical processing. This indicates that numerous correlations present in the raw data were merely spurious and could lead to misleading conclusions if analyzed without preprocessing. By revealing genuine relationships and removing noise, Data Preparation established a more reliable analytical foundation, thereby enabling actionable insights for both target groups.

3. HOW - How do we conduct this process? Why is Data Preparation the key?

3.1. Starting with the Big Picture (Data Understanding)

The analytical process begins by observing the overall landscape of the dataset, including the age distribution, the distribution between the two occupational groups (Students and Workers), and the corresponding depression rates. This is a necessary step to establish a foundational understanding of the data structure and to identify irregularities that may influence the entire analytical narrative. Beginning with these fundamental distributions ensures that subsequent conclusions are drawn within the correct context of the studied population.

At the same time, examining the raw data prior to cleaning plays a crucial role in detecting elements that can distort insights. Raw data allow the identification of instability indicators such as unreasonable age distributions, missing values in key variables, or abnormal entries within behavioral and emotional measures. These issues not only undermine the reliability of analytical models but can also lead to constructing an entirely inaccurate narrative about the depression landscape of the two groups. Therefore, understanding the raw data is considered an essential "scene inspection" before conducting an analytical "forensic investigation," ensuring that the overall picture is built upon a sound foundation from the beginning.

3.2. Data Cleaning

In this step, the focus lies not on the technical procedures of cleaning but on why cleaning is a prerequisite for ensuring that insights are not distorted. One common issue is a high proportion of missing values in sensitive variables such as depression scores or psychological pressure levels. If these values are not handled appropriately, overall depression rates may be skewed, leading to inaccurate assessments of the severity of the issue within each group.

In addition, the presence of outliers-for instance, cases reporting working or sleeping more than 24 hours a day-creates the illusion that overwork is an extreme factor strongly associated with depression. Retaining such unrealistic values leads the analytical model to exaggerate the influence of work-time-related factors, resulting in a biased narrative. Similarly, pressure scores that contain

substantial noise weaken or destabilize the correlation between pressure and depression, obscuring the truly relevant relationships among variables.

Data cleaning therefore functions as the stage of "setting up the scene," ensuring that subsequent Data Processing and advanced analysis can proceed coherently and accurately. When the data foundation is free of noise, structurally normalized, and redistributed appropriately, the analytical story can be constructed around genuine relationships rather than being guided by false or distorted signals from the outset.

3.3. Data Transformation & Feature Engineering

In the analytical process, creating new variables is not a matter of technical manipulation but a means of illuminating underlying patterns and narratives within the data. One key step involves constructing a stress index that consolidates multiple discrete factors into a single measure. Instead of examining each stress-related variable independently, this index provides a coherent picture of overall stress levels, thereby clarifying patterns that raw data cannot easily reveal. This is particularly important when comparing specific groups or cohorts with shared characteristics.

Similarly, age grouping helps identify high-risk stages more intuitively. By observing age bands rather than individual ages, the 18-25 range stands out as a period of heightened pressure, reflecting the transition from adolescence to adulthood with numerous personal and societal expectations. This insight only emerges once the data are restructured into meaningful age categories.

Furthermore, constructing a composite lifestyle score that integrates behavioral aspects such as sleep, diet, physical activity, and social interaction creates a coherent narrative linking lifestyle and mental health. Instead of analyzing each behavior separately, the composite score allows an assessment of cumulative influences and highlights individuals living under persistently adverse conditions. As a result, the narrative connecting lifestyle and depression becomes more consistent, communicable, and actionable.

Overall, creating composite variables not only clarifies relationships among factors but also ensures that the analytical story is built from data that are structured, interpretable, and intuitive.

3.4. Segmentation

Segmentation is not merely an analytical technique but a necessary choice to ensure that the narrative accurately reflects the lived realities of two groups whose life contexts differ entirely. Students and Workers enter adulthood through two distinct pathways, carrying different resources, pressures, and responsibilities. Therefore, analyzing them jointly would blur essential differences and obscure the unique mechanisms affecting each group.

For Students, stress is primarily driven by endogenous factors: academic expectations, peer pressure, social comparison, and financial dependence. Meanwhile, Workers face more exogenous pressures, such as job demands, family responsibilities, and long-term financial burdens. These two stress ecosystems result in two different mechanisms of depression-something that a combined analysis cannot reveal.

Segmentation therefore becomes a tool to "tell two stories within one report" clearly and with depth. Instead of forcing a single model onto all individuals, segmentation allows the presentation of two parallel adulthood trajectories and two psychological mechanisms without blending them. This not only

enhances the accuracy of insights but also ensures that subsequent recommendations can be tailored to the specific needs of each group.

3.5. Analysis & Narrative Construction through Visualizations

In the analytical stage, the selection of each visualization plays a vital role in guiding the reader through the logic of the data narrative. To establish the initial anchor points of the story, a heatmap is used to outline the overall pattern of relationships among variable groups. This is not a tool for final conclusions but a directional device that highlights noteworthy associations and areas requiring deeper investigation in later steps.

To illustrate the difference between raw and cleaned data, the boxplot is chosen for its ability to clearly depict outliers and distribution spreads. It allows readers to visually observe how abnormal values in the raw data distort distributions, thereby explaining why initial insights may be inaccurate. The shift in the boxplot's shape before and after cleaning reinforces the argument that Data Preparation is a prerequisite for trustworthy analysis.

Next, bar charts are employed to emphasize disparities between the two groups. This visualization is particularly effective for direct comparisons, illustrating differences in depression rates, stress levels, or lifestyle scores between Students and Workers. Bar charts function as visual highlights, enabling readers to quickly recognize the core distinctions central to the narrative.

Finally, scatter plots are used to expose how raw data can obscure relationships. By placing raw and cleaned data side by side, the scatter plot clearly shows how outliers, missing values, or noise previously distorted trend lines. After cleaning, relationships become more coherent, providing a reliable foundation for analytical conclusions. This underscores the point that a data story can only be told correctly when the underlying dataset has been standardized and properly processed.

3.6. Turning Insights into Actions

Data analysis does not end with describing the current situation; it must move toward actionable outcomes. Based on the findings validated through Data Preparation, mental health support recommendations can be designed specifically for each target group. For Students, programs may focus on academic pressure management, improving self-awareness skills, reducing social comparison, and enhancing financial support or scholarships to alleviate dependency. Meanwhile, Workers may benefit from programs that improve job satisfaction, manage workload, support work-family balance, and provide long-term financial counseling.

Additionally, mental health policies in educational institutions, workplaces, or community organizations can be optimized based on standardized data. Using reliable data enables the design of interventions that target the right groups, address the right causes, and allow impact measurement over time. This ensures that the report's recommendations are not merely theoretical but can be translated into concrete, evidence-based actions suited to the specific contexts of each population group.

III. DATA PREPARATION PIPELINE

1. Data processing

1.1. Overview

In this section, we present the full workflow of cleaning and standardizing the data before feeding it into the model for training. The original dataset contains many input errors, noisy data, and missing values, so it must be handled carefully to avoid model bias.

An overview of the original dataset, it contains 170,700 rows with 20 columns, with many missing values in variables such as Study Satisfaction, Academic Pressure, CGPA, Profession, Work Pressure, and Job Satisfaction, along with many columns containing corrupted text, typos, garbage values, and inconsistent data between the Student group and the Working Professional group.

1.2. Initial Cleaning

Before going into more complex data-cleaning steps, we decided to remove redundant columns: id and Name. Since these are only identifiers of the survey respondents, they have no correlation with mental health status (Depression), so removing them helps keep the dataset lighter and allows the model to focus on more meaningful features.

1.2.1. Handle Logic

In the dataset, the variable “Working Professional or Student” was self-reported by respondents. However, several inconsistent cases appear: some respondents selected Student but filled in columns that only apply to working individuals, and conversely — some selected Working Professional but filled in study-related information, including CGPA. These inconsistencies most likely come from mis-selection during the survey, leading to logical conflicts between variables and potentially degrading model quality.

Therefore, we decided to correct this logic based on actual data-entry behavior. Specifically:

- If a respondent selected Student but filled in more work-related columns, we changed their label to Working Professional.
- Conversely, if a respondent selected Working Professional but filled in more study-related information, they were changed to Student.

After updating the labels, any values appearing in columns that do not belong to their group (e.g., a Student with filled work-hours information) were set to NaN for processing in the next steps. This approach helps clearly determine who is a Student and who is a Working Professional, ensuring dataset consistency and supporting later preprocessing and modeling.

1.2.2. Likert Scale Variables

The variables Academic Pressure, Study Satisfaction, Work Pressure, Job Satisfaction, and Financial Stress contain textual data with many invalid values such as “Low”, “High”, “Error”, “Null”, or values not convertible to numbers. We standardized descriptive values like “Low” and “High” into

numeric form on a 1–5 Likert scale, converted all data into integers, and transformed invalid values including “Error”, “Null”, “??”, “#VALUE!”, etc., into NaN. After applying these steps, these variables contain only numerical values representing level 1 to 5.

1.2.3. Categorical Data Standardization

The City column contains special characters at the beginning (such as *Kalyan, -Surat), spelling mistakes (Molkata → Kolkata), and many garbage values (names of people, occupations, degrees). We standardized the text by removing stray characters using regex, correcting spelling, and converting invalid entries into NaN. The result retains 29 distinct cities.

The Gender column contains many inconsistent variants (uppercase/lowercase mixing, abbreviations like M, f, and synonyms like Boy, Woman). To avoid unnecessary category inflation, we standardized the values into only three groups: Male, Female, Other, and removed rows with invalid gender.

The Dietary Habits column also contains inconsistent values—mixed capitalization, numeric strings, and unusual labels. All values were standardized into only three valid categories: Unhealthy, Moderate, Healthy. All other invalid categories were converted to NaN.

For the Sleep Duration column, values were inconsistent (7-8, 6-8, More than 8), contained meaningless values such as 45, Pune, and spelling errors. We built a keyword-based mapping to group them into three standard categories: More than 8 hours, 6–8 hours, Less than or equal to 6 hours. Garbage values were converted into NaN and removed from the dataset.

The Degree column initially contained a large amount of noise: inconsistent abbreviations, mixed with occupations (Plumber, Manager), numerical strings, special characters, and other garbage values. To standardize, all values were first lowercased, then grouped into four major categories: Doctorate, Master, Bachelor, and High School. Values that do not belong to any education level were converted into NaN because they are not meaningful for this variable.

Similarly to Degree, the Profession variable suffers from multiple issues: misspellings, inconsistent grammar, noisy values such as names of people, cities, degrees, and many missing values. Therefore, the cleaning process had to be carried out in multiple steps:

- Standardizing text format: converting all values to title() for consistency.
- Correcting spelling variations of job titles to map them into valid occupation groups.
- Removing noisy values: any item not related to a profession (e.g., person names, locations, degrees) was converted into NaN for later imputation.

Due to the characteristics of this dataset, the Profession variable has a direct relationship with the variable Working Professional or Student. Therefore, imputation was performed separately for the two groups:

- For Students: missing Profession values were assigned “Student”. This is reasonable because most missing values belong to this group and the dataset only has two main states (Student vs. Working Professional), so no hybrid cases exist.
- For Working Professionals: individuals labeled as Working Professional who reported their profession as “Student” or left it blank were assigned “Unspecified” to maintain logical consistency.

Then, we grouped the professions into Student, Education & Research, Medical & Health, Tech & IT, Business & Finance, Creative & Media, Engineering & Architecture, Legal, Service & Operations, and Other Professions for later processing.

1.2.4. Physical Consistency & Range Constraints

Some numerical columns contain unrealistic values inconsistent with physical or real-world constraints, so we applied range restrictions and initial standardization as follows:

- CGPA: Some values were abnormal (e.g., 100). We restricted valid CGPA values to 0–10; any value outside this range was set to NaN.
- Age: Values less than 0 or greater than 100 appear. We restricted the valid age to 10–90; values outside this range were converted to NaN.
- Work/Study Hours: Some values were less than 0 or greater than 24. Any value outside the 0–24 range was set to NaN. Note: these NaN values were not removed because these columns contain important information for prediction — they were imputed in the next steps.

After applying these checks and converting invalid values to NaN, the data became consistent in basic logic and formatting, ready for splitting and further preprocessing.

1.3. Split Data

Since the goal is to build models tailored to each demographic group — as students and working professionals have different psychological characteristics and habits that may affect depression levels — the data was split into two separate groups based on the variable “Working Professional or Student”. Each group was split into Train/Test with an 80/20 ratio. The stratify parameter was used to preserve the proportion of the “Depression” label in both sets, ensuring that the model is not distribution-biased. Splitting the dataset helps improve generalization and increases model accuracy for each subgroup.

1.4. Handle Missing Value

1.4.1. Imputation for Numeric Variables

For study-related, work-related, and time-related columns, we imputed the data separately for Student and Working Professional groups to ensure the data is handled according to the true behavioral characteristics.

In the Student group, the variable CGPA contains some missing values. Since this is a continuous variable with strong relationships with academic-related features such as Academic Pressure, Study

Satisfaction, and Age, mean/median imputation is not suitable. Instead, we used KNN Imputer, because this method estimates missing values based on the closest instances, preserving natural relationships among features. The KNN model was fitted on Train and applied to Test only to avoid data leakage. In the Working Professional group, CGPA is irrelevant and was assigned 0 for consistency.

For academic and work-related variables, the data was processed separately:

- Student: only study-related variables (Academic Pressure, Study Satisfaction, CGPA, ..) were imputed using KNN; work-related variables were assigned 0.
- Working Professional: only work-related features (Work Pressure, Job Satisfaction..) were imputed via KNN; study-related variables were assigned 0.

This approach ensures that each group is imputed based on its behavioral patterns and avoids data leakage between Train/Test.

Similarly, the Work/Study Hours variable was imputed with KNN after invalid values were converted to NaN. Each group used relevant features — Age, academic/work pressure, satisfaction levels, Financial Stress, etc. — so the model could predict missing values accurately, while unrelated features were set to 0 to avoid noise.

1.4.2. Imputation for Categorical Variables

The variable Dietary Habits is categorical, so mean/median imputation is impossible. Also, dietary habits do not strongly correlate with academic or work-related variables, so using complex methods like KNN is unnecessary. Therefore, we used Mode Imputation separately for Students and Working Professionals.

For each group, the mode was computed on Train and then applied to both Train and Test to avoid data leakage and preserve natural differences in the distribution of eating habits between groups. This ensures consistent data while reflecting true behavioral characteristics of each demographic in the dataset.

1.5. Outliers

To ensure a stable distribution of Age and reduce the influence of extreme values, the Age variable was processed using IQR Capping, performed after splitting Train/Test to avoid data leakage. On the Train set of each group (Student and Working Professional), we computed Q1, Q3, and IQR, then determined the bounds:

- Lower bound = $Q1 - 1.5 \times IQR$
- Upper bound = $Q3 + 1.5 \times IQR$.

These thresholds were rounded to the nearest integer to match the discrete nature of age. All values outside the safe range were capped to the corresponding threshold in both Train and Test. This ensures that the distribution of Age remains clean, without distorting the original shape, and remains consistent across datasets.

The preprocessing pipeline has thoroughly addressed the major issues of the original dataset: from standardizing messy categorical variables (City, Degree, Profession) to logically restoring missing values using KNN. Most importantly, splitting and processing the data separately for Students and Working Professionals ensures that the data accurately reflects each group's behavioral characteristics and eliminates all logical inconsistencies. As a result, we obtain a clean dataset without extreme outliers and completely free of noise, forming a strong foundation for building high-accuracy machine-learning models in the next steps.

2. Feature Engineering

The feature engineering pipeline was designed to standardize the informational structure of the dataset and to ensure optimal learnability for machine learning models. After data cleaning, the dataset was separated into two independent branches: Students and Working Professionals, to accurately reflect the inherent differences between these two subpopulations.

2.1. Data Transformation

2.1.1. Categorical Encoding

Categorical variables were encoded using three different mechanisms depending on the statistical nature and structural properties of each variable group.

- For non-ordinal categorical variables, including Gender, Profession, and Degree, One-Hot Encoding was applied to eliminate any artificial assumptions of linearity or hierarchical ordering among categories. Gender consists of three labels ("Male", "Female", "Other"), while Profession includes ten unique categories and Degree comprises seven distinct educational levels. Although Degree categories may appear to follow an educational hierarchy, the underlying dataset does not support a consistent ordinal structure or any interpretable quantitative spacing between these labels; therefore, One-Hot Encoding is the most appropriate approach to avoid misleading ordinal implications.
- For variables with an inherent ordered structure, namely Sleep Duration and Dietary Habits, Ordinal Encoding was employed to preserve their internal hierarchy. Sleep Duration is categorized into "≤6 hours", "6–8 hours", and ">8 hours", reflecting ascending levels of rest. Similarly, Dietary Habits comprises "Unhealthy", "Moderate", and "Healthy", capturing increasing degrees of nutritional adequacy. Maintaining ordinal integrity is crucial, as the ordered progression carries meaningful implications for physical and mental well-being.
- For high-cardinality variables with imbalanced category distributions, such as City, Frequency Encoding was adopted. This method reduces dimensionality while preserving the statistical signal associated with the relative prevalence of each category, thereby maintaining essential contextual information related to living conditions, socio-economic environment, or demographic patterns that may influence mental health outcomes.

2.1.2. Creation of New Features and Interaction Terms

To enable the model to capture complex psychosocial mechanisms that cannot be expressed by single variables, multiple engineered features and interaction terms were constructed.

a. Feature Interactions for Student Data

For students, interaction terms primarily focused on relationships among academic pressure, study workload, satisfaction levels, and academic performance.

- Burnout_Load = Academic Pressure × Work/Study Hours: models the risk of academic overload. Students experiencing high academic pressure and long study hours are more prone to academic burnout. Conversely, high pressure combined with low study hours may indicate avoidance coping or lack of motivation, often associated with external stressors such as parental expectations or performance anxiety.
- Effort_Reward_Imbalance = Burnout_Load × (6 – Study Satisfaction): represents the imbalance between effort (pressure × workload) and subjective reward (satisfaction). A large imbalance indicates that substantial effort is not matched with adequate satisfaction, a pattern strongly linked to heightened depression risk.
- Performance_Anxiety_Score = CGPA × Academic Pressure: captures the interaction between academic achievement and pressure. Low GPA combined with high pressure reflects vulnerability to academic anxiety and depressive symptoms. High GPA with high pressure reflects the "perfectionist" profile, well-documented in literature as a group susceptible to heightened stress due to overly rigid self-imposed standards.

b. Feature Interactions for Working Professionals

For the working group, interaction features were developed to capture occupational stress, effort–reward discrepancies, and financial or motivational constraints.

- Burnout_Load = Work Pressure × Work/Study Hours: This models the risk of workplace burnout among individuals subjected to long working hours under high pressure.
- Effort_Reward_Imbalance = Burnout_Load × (6 – Job Satisfaction): Individuals who expend substantial effort but experience low job satisfaction often suffer from emotional exhaustion, loss of motivation, and psychological strain.
- Trapped_Score = Financial Stress + (6 – Job Satisfaction): represents occupational entrapment, where financial constraints combined with low job satisfaction may prevent individuals from leaving their job, leading to prolonged psychological distress.

c. Common Interaction Features for All Participants

A set of cross-group interaction features was constructed to capture shared determinants of mental health, including physical well-being, sleep patterns, financial stress, and overall sensitivity to psychosocial pressure.

- Physical_Health = Sleep Hours × Dietary Habits: models the combined effect of sleep and nutrition on physical health which are two well-established predictors of mental well-being.

- Work_Sleep = Work/Study Hours × Sleep Hours: quantifies the balance (or imbalance) between workload and rest. High workload and limited sleep are known predictors of burnout and reduced psychological resilience.
- Suicidal_Family_Interaction = Suicidal Thoughts × Family Mental History: captures the compounded risk associated with both individual suicidal ideation and familial mental health predispositions.
- Sleep_Vulnerability = Financial Stress × Sleep Hours: Financial strain is known to disrupt sleep quality. This interaction reflects heightened vulnerability to depressive symptoms arising from stress-induced sleep disturbances.
- Financial_Work_Interaction = Financial Stress × Work/Study Hours: quantifies compensatory behaviors such as working or studying excessively due to financial pressure, potentially leading to exhaustion and poorer mental health.
- Age_Pressure = Pressure / Age: models the empirical observation that younger individuals tend to experience stronger psychological reactions to equivalent levels of stress compared to older individuals.

2.1.3. Numerical Scaling

Numeric features were standardized using StandardScaler, which transforms each variable to a distribution with mean 0 and standard deviation 1. Despite the fact that RandomForestClassifier does not inherently require feature scaling, scaling was still implemented for methodological consistency across the entire preprocessing pipeline and to prevent variables with large magnitudes from disproportionately influencing feature importance estimates. The standardized representation also facilitates model comparability and improves numerical stability in downstream analytical procedures.

2.2. Feature Selection

Given the moderately high number of predictors, including both original and engineered features, feature selection was conducted to ensure that the predictive model concentrates on the most informative signals.

- Pearson correlation was used to quantify linear associations between each predictor and depression severity, enabling identification of variables with direct linear influence.
- Mutual Information was employed to detect non-linear dependencies that Pearson correlation cannot capture.
- Variance Inflation Factor (VIF) was used to diagnose multicollinearity among predictors, ensuring removal of redundant features that compromise model interpretability.
- Embedded selection methods, including LassoCV and feature importance scores from Random Forest, were used to evaluate each feature's contribution within actual learning models.

The final feature set was determined using an aggregate ranking approach, whereby the results from all selection methods were combined into a mean-rank score. From this, the top 10 most relevant features were selected for visualization and for inclusion in the final predictive model.

IV. THE PRINCIPLES OF VISUALIZATION

This section presents the theoretical framework that guides the entire visualization design for the Data Storytelling part of our project on factors influencing depression among students and working professionals. The overarching goal is to build a visualization system that is consistent, intuitive, and scientifically grounded. Every chart is designed to help the reader clearly see the difference between raw and cleaned data, and understand which factors exert the strongest influence on depression.

The section includes four components:

(1) the rationale for using a blue color palette and its meaning in the mental-health context;

(2) the application of pre-attentive attributes to guide visual attention effectively;

(3) the principles of removing visual clutter to reduce cognitive load;

(4) the rationale behind choosing the appropriate chart types.

1. Meaning of the Blue Color Palette in the Context of Depression

Blue was selected as the primary palette due to its strong alignment with both color psychology and the emotional nature of the depression topic. Psychologically, blue is associated with calmness, emotional stability, and trust - qualities that make it suitable for visualizing sensitive mental-health variables such as stress, burnout, and depression. In contrast to warmer colors like red or orange, which create visual tension, blue is softer, more grounded, and less stimulating, helping the audience absorb emotionally heavy information comfortably.

Using a consistent blue palette also creates visual harmony across the report, preventing distraction caused by switching among multiple hues. Blue carries a dual emotional meaning that fits the theme: the “feeling blue” metaphor reflects empathy for psychological struggles, while softer and brighter tones of blue symbolize hope, recovery, and balance. This combination enables the charts to maintain a serious, empathetic tone without dramatizing the issue.

2. Use of Pre-Attentive Attributes

In this project, pre-attentive attributes serve as a foundation for guiding the audience’s attention toward the key signals in the data - particularly variables related to stress, academic and work pressure, and depression symptoms. The intention is not to explain the theory itself, but to ensure that every chart communicates the main idea effortlessly, reduces cognitive overload, and aligns with the storytelling goals of the report. Choices were made based on the structure of the dataset (quantitative, categorical, Likert scale variables) and the insights we needed to reveal.

2.1. Attributes for Quantitative Data

Quantitative variables such as stress levels (1- 5), pressure, satisfaction, or depression scores benefit most from pre-attentive attributes that highlight magnitude differences quickly.

2.1.1. Position

Position on a common axis - used in charts - is the most powerful pre-attentive attribute. For stress levels or depression rate shifts, position makes increasing or decreasing trends immediately visible, without requiring additional visual elements.

2.1.2. Length

Bar charts are used when direct comparison across groups is needed (e.g., Student vs Working Professional). Height differences can be recognized instantly, making it ideal for storytelling moments where we emphasize “who is more at risk?”

2.1.3. Color Intensity (Blue Gradient: Light → Dark)

A controlled blue gradient encodes magnitude:

- Light blue → low level
- Medium blue → moderate level
- Dark blue → high level (critical)

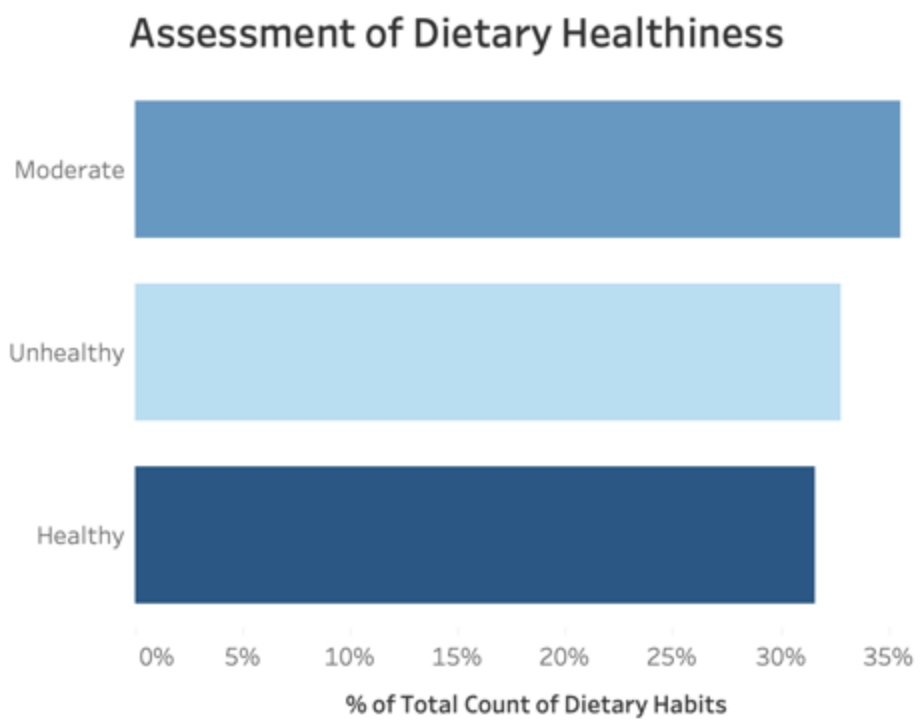
This approach helps the viewer identify areas of high stress or pressure at a glance and is especially effective for variables measured on ordered scales (e.g., Academic Pressure, Work Pressure).

2.2. Attributes for Categorical Data

Categorical variables (Gender, Profession, City, Sleep Durations, Dietary Habits) require distinct categories without implying order.

2.2.1. Hue

A limited set of blue-based hues is used to separate categories such as Students vs Working Professionals or Healthy vs Moderate vs Unhealthy eating habits. Using hues within the same color family preserves aesthetic consistency and avoids overstimulation—important for a mental-health topic. Minimal use of hue keeps charts clean and focused.



3. Principles of Removing Visual Clutter

Throughout the visualization system, the "Remove Clutter" principle ensures that readers stay focused on the insights—namely, the factors contributing to depression among students and workers. Instead of adding decorative elements or excessive details, we prioritize clarity and simplicity so the core message is understood at first glance.

3.1. Removing Excess Gridlines and Borders

All heavy or unnecessary gridlines are removed because they add little value to comparison tasks. Removing borders also creates a more open composition, allowing the data to stand out.

3.2. Reducing Tick Marks and Labels

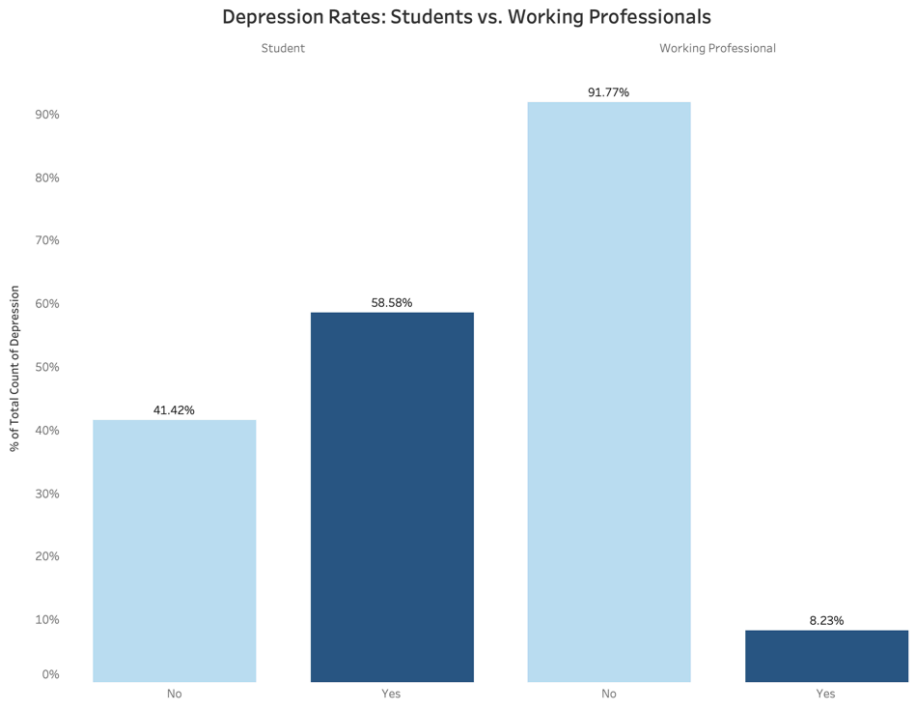
Axis ticks are limited to essential reference points. Overly long labels or repetitive descriptions are simplified. This prevents the reader from being visually overloaded by numbers.

3.3. Limiting the Number of Colors

A strict rule in our design system is to avoid using too many colors in a single chart. A uniform blue palette prevents attention from being diverted to irrelevant colors and supports the calm tone of the topic. Even for binary categories like Yes/No, only two subtle blue tones are used—avoiding the red–green or red–yellow combinations common in dashboards, which can feel harsh or emotionally charged.

3.4. Removing Unnecessary Legends

Whenever possible, legends are removed because the information is already conveyed through labels or titles. This elimination decreases eye movement and makes charts easier to interpret.



Through these principles, the entire visualization set becomes clean, coherent, and aligned with the emotional and scientific nature of the topics.

4. Choosing the Right Chart Types

Every chart type was selected using a "story-first" approach: each visualization must help explain part of the narrative about how various factors influence depression, while simultaneously highlighting the contrast between raw and cleaned data.

4.1. Bar Charts / Countplots – Highlighting Group Differences

Bar charts are essential for categorical comparisons such as Depression Yes/No or levels of Work and Academic Pressure. Length is a strong pre-attentive cue, making bar charts ideal for quickly revealing imbalances. These charts support the narrative by showing the overall landscape before diving deeper.

4.2. Histograms / KDE Plots – Showing Distribution Shapes (Raw vs Clean)

Histograms and KDE plots show the underlying distribution of quantitative variables like Age. They make data abnormalities visible—skewness, outliers, irregular clusters—which helps explain why preprocessing is necessary. Overlaying raw and clean distributions on the same axis allows the viewer to see the effect of preprocessing without lengthy explanation.

4.3. Boxplots – Highlighting Variability and Outliers

Boxplots emphasize distribution spread and outliers, which are crucial in mental-health datasets where variations often signal instability or extreme conditions. For variables like Age or Work/Study Hours, boxplots reveal differences not only in central tendencies but also in volatility. They clearly show the reduction of noise after cleaning, making them powerful for before-and-after comparisons.

4.4. Scatter Plots – Showing Relationships and Noise Levels

Scatter plots illustrate relationships between variables such as Burnout Load and Average Depression. They help the viewer see both the underlying trend and the degree of noise. Raw data often appears as a dense cloud, while cleaned data reveals a clearer pattern. This visual contrast is essential for demonstrating how preprocessing affects interpretation.

5. Simplified Design

All charts follow a simplified 2D design with no shadows, textures, or 3D effects. Each chart conveys a single message and uses at most one or two highlighting attributes (color and position). This minimalistic approach ensures clarity, maintains emotional sensitivity given the mental-health context, and keeps attention focused on the difference between raw and clean data.

6. Conclusion

The visualization system in this project is built on a coherent set of principles—from color selection and pre-attentive attributes to clutter reduction and chart-type choices. As a result, every visualization not only presents data but also helps tell the story of how preprocessing transforms insights and reveals the true factors contributing to depression. The approach ensures clarity, scientific rigor, and emotional appropriateness for such a sensitive topic.

V. OVERVIEW AND METHODOLOGY IN DEPRESSION RISK PREDICTION USING MACHINE LEARNING

1. Research Objectives and Scope

This study aims to develop and evaluate the effectiveness of a machine learning–based classification system designed to predict the risk of depression. The core focus of the analysis lies in conducting a controlled comparison between the performance of a model using original features (Baseline) and a model integrating advanced engineered features (Enhanced). To ensure accuracy and alignment with psychosocial characteristics, the study applies a stratified subject segmentation strategy from the early stage of data preprocessing.

2. Establishing the Baseline Model

To establish a standard benchmark, the baseline model is constructed to measure predictive capability when relying solely on raw information from cleaned data without applying complex feature transformation techniques. Recognizing the fundamental differences in living environments and psychological pressures across demographic groups, the study implements a Stratified Strategy at the data input stage. Specifically, the dataset is divided into two independent subsets: one for students and another for working adults.

Regarding the feature space, the baseline model operates entirely on the original variables available in the dataset, such as age, gender, academic pressure, and work pressure, without employing interaction terms or advanced feature selection techniques. The Random Forest Classifier (RFC) algorithm is selected and trained separately for each subgroup. However, despite the group segmentation, the baseline model still exhibits inherent limitations in semantic depth. The discrete handling of variables such as pressure and working hours prevents the model from capturing nonlinear synergistic relationships, for instance, the cumulative impact of high pressure sustained over long periods. Consequently, the accuracy of this model often reaches a performance ceiling due to being constrained by the surface-level information of the original features.

3. Enhanced Model Architecture and Feature Engineering Techniques

To overcome the limitations of the baseline model, the next stage of the study focuses on data enrichment through feature engineering. Building on the already segmented data, the enhanced model incorporates derived variables that offer deeper quantitative insights into psychological states.

Interaction terms play a central role in this improvement. Specifically, the variable Burnout_Load is established to represent accumulated exhaustion burden, calculated as the product of pressure and time, enabling the model to more precisely distinguish high-risk cases. The variable Age_Pressure reflects the proportional impact of pressure across age groups, supporting the identification of psychological sensitivity typical among younger individuals. Additionally, the Effort_Reward_Imbalance index measures the asymmetry between effort exerted and the resulting satisfaction or academic performance (CGPA), providing a multidimensional perspective on study and work motivation. To optimize inputs and prevent overfitting, the study employs Lasso Regularization combined with Mutual Information to select the 10 most informative features while eliminating noise from less important variables.

4. Experimental Results and Discussion

Experimental results indicate a substantial improvement in predictive performance when applying the enhanced model. For the student dataset, the overall accuracy reaches approximately 87%. More importantly, evaluation metrics for the “Depressed” class record a Precision of 0.88 and Recall of 0.90, yielding an F1-Score of 0.89. This improvement demonstrates that, thanks to derived variables such as Age_Pressure, the model more effectively identifies subtle depressive indicators that the baseline model typically overlooks, with the high Recall reflecting the system’s superior sensitivity.

For the working adult dataset, the model achieves an impressive accuracy of 96%, with the F1-Score exceeding 0.95. This outcome confirms the pivotal role of interaction variables, particularly Effort_Reward_Imbalance. In professional working environments, dissatisfaction stemming from an imbalance between effort and outcomes is often a primary driver of depression, and quantifying this factor enables exceptionally precise classification.

5. Conclusion

Overall, the substantial performance gains in the enhanced stage validate the study’s two main hypotheses. First, stratifying subjects is a necessary condition for isolating distinct psychological characteristics. Second, feature engineering serves as both a sufficient and essential condition. Transforming raw data into derived variables such as Burnout_Load and other interaction terms is the key to enabling the model to surpass the performance limits of traditional methods, turning discrete data into highly predictive and reliable knowledge.

VI. CONCLUSION

The project unequivocally confirms a core principle in modern Data Science: Data Preparation is not an optional step, but a mandatory foundation for achieving trustworthy and highly interpretable Predictive Models (ML Models). Success in identifying the factors influencing depression across the Student and Worker segments directly hinges on the systematic process of transforming subjective and noisy raw data into a structured, clean, and statistically aligned dataset suitable for model assumptions.

1. The Decisive Role of Data Processing in the ML Project

The Cleaning phase was critical in removing "surprising" correlations that only existed in the raw data. By eliminating noise and outliers, we prevented the ML model from learning misleading relationships, thereby ensuring the stability and genuine predictive accuracy of the final results.

Feature Engineering techniques, through the construction of composite indices like the stress index and composite lifestyle score, successfully illuminated complex patterns that the raw data obscured. This allowed the ML model to generate clearer actionable insights, specifically in distinguishing the underlying mechanisms of depression between Students (driven primarily by internal/endogenous factors) and Workers (driven primarily by external/exogenous factors).

The comparative use of charts like Boxplot and Scatter Plot before and after Data Preparation provided visual evidence. This visually demonstrated that the cleaned data resulted in clearer, more coherent variable distributions and relationships, validating the quality of the input and strengthening confidence in the model's predictive output.

2. Key Takeaway (Big Idea)

The analysis results strongly re-affirm the project's Big Idea: "When data is properly standardized, both insights and ML model performance significantly improve."

3. Limitations

The primary limitations of this study are rooted in the nature of the data and the chosen methodology. Firstly, the reliance on Self-Reported Data (Self-Reporting Bias) means key variables (e.g., Satisfaction Level, Pressure) are based on subjective perceptions. While rigorous Data Preparation addressed noise and outliers, this inherent subjectivity places a ceiling on data quality, which even systematic processing cannot fully overcome. Secondly, the Cross-Sectional Research Design is a fundamental limitation of the data structure itself, allowing only for the identification of correlations at a specific point in time and preventing the establishment of true causal links—a flaw that necessitates future longitudinal data collection rather than further data processing. Finally, concerning the algorithm, the choice of the Random Forest model, while prioritized for interpretability, represents a trade-off in Algorithmic Scope and Complexity. This model architecture may lack the capacity to fully capture the complex, higher-order non-linear feature interactions within psychosocial data, suggesting that exploring more sophisticated methods like advanced Gradient Boosting (XGBoost, LightGBM) could yield superior predictive performance metrics (e.g., AUC-ROC, F1-score).

4. Recommendations and Future Work

The study's findings, validated through a rigorous Data Preparation process, lead to targeted intervention recommendations. For Student Support Programs, proposals should prioritize resources offering Trauma-Informed Care, aimed at addressing residual psychological trauma and self-perception issues, which were identified as the strongest internal drivers of depression after data cleansing. Furthermore, implementing targeted programs for Financial Literacy & Dependency Reduction is crucial to mitigate the stress associated with financial dependency and social comparison, factors which the data clearly showed exacerbate stress levels within the Student segment. Conversely, for Employee Wellness Programs, businesses should concentrate on Stress & Satisfaction Alignment by addressing low job satisfaction and age-related pressure (comparison with peers), which were found to be statistically significant factors in the Worker segment. Finally, providing resources for Long-Term Financial and Family Support is necessary, as these external pressures were identified as critical, sustained stressors for Workers.

In addition, future research should build upon the established framework of data quality and predictive modeling. Specifically, it should focus on Advanced Feature Optimization & Selection: applying advanced explainability techniques such as SHAP/LIME on the cleaned dataset and the ML model to precisely quantify the contribution of each engineered feature (e.g., stress index, lifestyle score). This will provide a more granular understanding of feature importance, moving beyond basic correlational analysis. In parallel, exploring Automated Feature Engineering (AutoFE) tools can help discover higher-order, non-linear feature interactions missed by manual methods, potentially increasing the model's predictive ceiling (metrics like AUC-ROC, F1-Score). Methodologically, a shift from the current cross-sectional design to a Longitudinal Causal Inference Study is warranted to establish true causal links—for instance, to definitively determine whether low job satisfaction causes depression or is merely a symptom—thereby strengthening the evidence base for policy recommendations. Lastly, ensuring Model Generalization and Robustness involves validating the existing ML model on external, distinct datasets (e.g., in the Vietnamese or other cultural contexts) to test the robustness and generalizability of the learned feature weights. Concurrently, evaluating the performance gains achieved by deploying the cleaned dataset on more complex ML architectures (such as XGBoost, LightGBM, or simple Neural Networks) will help compare the trade-off between predictive accuracy and model interpretability.