# **Title:** Data Analysis Report on Anxiety and Depression Dataset

## 1. Overview

Mental health issues such as anxiety and depression have become increasingly prevalent in today's fast-paced and demanding society. The growing concern for psychological well-being has led to a surge in data collection and analytical efforts to better understand the contributing factors to mental health disorders. This project presents an in-depth data analysis on a dataset related to anxiety and depression. Through data cleaning, exploratory data analysis (EDA), visualization, and statistical testing, the project aims to uncover meaningful patterns, highlight key variables associated with mental health indicators, and provide insights that may inform prevention and intervention strategies.

## 2. Project Description and Objectives

The dataset used in this project, presumably collected via surveys or observational studies, contains information on various demographic, behavioural, and psychological factors. The primary objective of the analysis is to explore the relationships between these factors and individuals' depression scores. Specific goals include:

- Import and prepare the dataset for analysis
- Conduct exploratory data analysis to understand data structure and distribution
- Identify and handle missing values

- Generate descriptive statistics for numerical and categorical variables
- Visualize the data to detect patterns and potential outliers
- Analyse correlations among numeric features
- Explore the influence of categorical variables on depression scores
- Conduct basic statistical testing to assess associations

## 3. Methodology and Tools Used

The analysis was conducted using Python programming language within a Jupyter Notebook environment. The key libraries and tools used include:

- **pandas:** for data manipulation and cleaning
- **matplotlib and seaborn:** for creating static, animated, and interactive visualizations
- **missingno:** to visualize missing data patterns
- **scipy.stats:** for conducting statistical tests such as Chi-Square

**Steps followed:**

- **Data Loading:** The dataset was imported using pandas, and its structure was explored using df.info() and df.head().
- **Missing Value Analysis:** The extent of missing data was analyzed both numerically and visually using missingno.matrix() and missingno.heatmap().
- **Descriptive Statistics:** Summary statistics were computed to understand central tendencies and variability.

- **Unique Value Counts:** A count of unique values for each feature was generated, which is useful for identifying categorical variables.
- **Distribution and Outlier Analysis:** Histograms and boxplots were created for each numerical column to study distributions and detect outliers.
- **Correlation Heatmap:** A heatmap of Pearson correlation coefficients was created to visualize relationships among numerical variables.
- **Target Feature Analysis:** Depression scores were treated as the target variable. Relationships between depression scores and other features were visualized using scatter plots and grouped bar charts.
- **Categorical Variable Analysis:** Bar plots were used to show how depression scores vary across categories, and

Chi-Square tests were conducted for associations between categorical variables.

**4. Key Tasks and Contributions**

The project included several core tasks that were successfully completed:

- **Data Cleaning:** Identified missing values and explored their patterns to determine data quality.
- **Exploratory Data Analysis:** Uncovered key trends and insights by visually and statistically analyzing data distributions.
- **Visualization:** Used multiple types of plots (histograms, boxplots, scatter plots, heatmaps) to summarize data and identify patterns.

- **Feature Analysis:** Investigated both numerical and categorical features and their impact on depression scores.
- **Preliminary Statistical Inference:** Performed Chi-Square tests to explore dependencies between categorical features.

These contributions demonstrate a comprehensive approach to understanding and presenting data in a meaningful way. The student successfully applied core data science skills such as preprocessing, EDA, and hypothesis testing.

**5. Results and Insights**

Some of the key insights derived from the analysis include:

- Certain features showed strong correlations with depression scores, which may indicate their potential as predictive variables.
- Distributions of numerical variables were examined, and some were found to be skewed, suggesting the need for normalization in future modelling.
- Boxplots revealed the presence of outliers in some features, which could impact statistical modelling.
- Grouped bar charts suggested that demographic or lifestyle variables might influence depression scores.
- Chi-Square tests highlighted statistically significant associations between some categorical variables.

These findings provide an initial foundation for building predictive models or designing targeted interventions.

**6. Conclusion and Future Work**

This project provides a strong foundational analysis of a mental health dataset focused on anxiety and depression. The comprehensive exploratory analysis offers valuable insights into

potential risk factors and relationships. In future work, the analysis can be extended in the following directions:

- Implement machine learning models to predict depression scores or classify individuals at risk
- Perform feature selection and engineering for better model performance
- Apply advanced statistical methods or deep learning approaches
- Explore temporal or longitudinal aspects if data is available

Overall, this project successfully demonstrates how data analytics can be used to gain actionable insights into mental health challenges and lays the groundwork for future predictive or prescriptive models. These may include supervised learning algorithms such as linear regression, decision trees, or ensemble methods like Random Forests and Gradient Boosting Machines to predict depression scores. Additionally, unsupervised learning techniques, such as clustering (e.g., K-Means or hierarchical clustering), could help group individuals based on similar behavioural or psychological profiles. Time-series analysis and longitudinal modelling can be pursued if the dataset includes temporal information. Furthermore, applying techniques like Principal Component Analysis (PCA) for dimensionality reduction or using Lasso/Ridge regression for feature regularization may enhance model interpretability and performance. Integration with Natural Language Processing (NLP) approaches can be explored if textual responses are available in the dataset.