Published by :
https://www.ijert.org/
An International Peer-Reviewed Journal

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

The Impact of Lifestyle Habits on the Mental Health of Adults

Yap Zi Bin
Group 8 of Statistical Inference and Modeling
Bachelor of Computer Science (Honours)
Taylor's University Lakeside Campus
Subang Jaya, Malaysia

Yip Qian Ling
Group 8 of Statistical Inference and Modeling
Bachelor of Computer Science (Honours)
Taylor's University Lakeside Campus
Subang Jaya, Malaysia

Yap Hui Ee
Group 8 of Statistical Inference and Modeling
Bachelor of Computer Science (Honours)
Taylor's University Lakeside Campus
Subang Jaya, Malaysia

Toh En Suen
Group 8 of Statistical Inference and Modeling
Bachelor of Computer Science (Honours)
Taylor's University Lakeside Campus
Subang Jaya, Malaysia

Chew Ying En
Group 8 of Statistical Inference and Modeling
Bachelor of Computer Science (Honours)
Taylor's University Lakeside Campus
Subang Jaya, Malaysia

Chew Ying Tong
Group 8 of Statistical Inference and Modeling
Bachelor of Computer Science (Honours)
Taylor's University Lakeside Campus
Subang Jaya, Malaysia

Abstract*—Mental health is a major factor affecting overall well-being and quality of life. With research showing that lifestyle-related behaviors such as diet, sleep, work hours, exercise and social interaction play an important role in shaping psychological well-being. With that being said, mental health concerns have become increasingly popular worldwide. Although existing research highlights the influence of these factors, it typically examines them individually rather than collectively. Thus, there is a need to understand how multiple lifestyle behaviours interact to affect mental health outcomes. This study aims to investigate the combined impacts of multiple lifestyle habits on the mental health of adults with the measurement of self-reported happiness scores. The goal is to identify which lifestyle factors Using the Mental health and Lifestyle Habits (2019-2024) dataset from Kaggle which contained 3000 individuals, this study applies a structured data preprocessing steps including encoding variables, handling outlier, scaling and feature engineering to prepare the data for further analysis. Besides, statistical models such as one way ANOVA, Pearson correlation and sample t-tests will be conducted to determine the relationships between lifestyle behaviours and happiness scores. The expectation of this study is to find whether healthier lifestyle habits like regular exercise, balanced diet, strong social interaction and adequate sleep are associated with happiness scores. In conclusion, the results indicate no strong significant relationships between the lifestyle habits and mental health outcomes, with most analyses failing to reject the null hypothesis. The findings emphasize the complexity of mental health outcomes and suggest that adult happiness may not be strongly explained by individual lifestyle habits alone. As a result, this reflects the need for more non-linear approaches and comprehensiveness in future mental health research.

Keywords—mental health, lifestyle habits, happiness score, data preprocessing, statistical analysis, regression analysis

I. INTRODUCTION

Nowadays, many adults experience increasing levels of stress due to work demands, lifestyle changes and social pressures. Mental health plays an important role in overall physical well-being which can influence an individual's emotional stability and quality of life. Lifestyle habits such as sleep hours, exercise, screen time, work hours, diet type and social interaction are part of daily life that might influence an adult's happiness level. In this study, the happiness score which is the mental health outcomes is used to evaluate by combining the factors that mentioned above. The purpose of this study can provide an insight into the overall psychological state of the individuals.

II. PROBLEM STATEMENT

Mental health is one of the important factors that influence an individual's overall well-being and quality of life. There are many factors that influence an adult's mental health, especially lifestyle habits. For instance, daily routines like sleep duration, working hours and diet quality can directly affect emotional stability and stress levels. However, despite growing awareness, limited studies have examined how these lifestyles collectively influence the mental well-being of an adult. Furthermore, the term collectively highlights the overall impact of all lifestyle factors acting together, rather than focusing on one or two factors in isolation. Therefore, this report aims to address the research gap by investigating the relationships between lifestyle habits as stated in the research question and mental health outcomes, which are measured through happiness scores among adults. In this study, the happiness score is considered the mental health outcomes, as higher scores generally indicate better psychological well-being and lower risk of mental distress. However, it may not capture all dimensions of mental health such as anxiety or depression severity.

Research Question:

Is there a statistically significant relationship between adults' lifestyle habits, which is the combination of factors such as exercise level, diet type, sleep hours, work hours per

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

week, screen time per day, and social interaction score, with their mental health outcomes (happiness scores)?

The outcomes of this study are expected to provide a better understanding of how daily lifestyle habits impact mental health and to promote greater awareness of mental well-being among adults. Hence, individuals can take proactive actions to improve their psychological health based on the insights gained from this study. To achieve this, this study utilizes the Mental Health and Lifestyle Habits dataset which contains comprehensive information such as diet type, exercise level and work hours on adult lifestyle behaviours and mental health conditions [1].

III. Research Objectives

There are four main objectives in our research paper, which aligns with our research questions, problem statement and the goals we needed to be achieved.

Data Analysis Phase	Research Questions	Research Objectives
Exploratory	How do multiple combinations of lifestyle habits influence the distribution and variation of mental health conditions among adults?	To identify the relationship between the multiple combination of lifestyle habits that can impact the mental health condition of adults by performing Exploratory Data Analysis (EDA) to analyze, and investigate the dataset patterns and distributions by using descriptive statistics and data visualization such as boxplots and correlations, to study the relationship and correlation between the multiple combination of lifestyle factors and the mental health condition of adults.
Modeling	Which combinations of lifestyle factors demonstrate a statistically significant relationship with the mental health condition of adults?	To develop efficient, analyzable, and visualizable statistical models to investigate the statistically significant relationship between multiple combinations of lifestyle factors with the mental health condition of adults based on the dataset. Therefore, perform data preprocessing and data cleaning for better visualization and analyzation of data.
Predictive	How to accurately predict an adult's mental health condition based on significant multiple combinations of lifestyle factors?	To build a predictive model for predicting the mental health condition outcome of adults based on the significant multiple combination of lifestyle factors. Therefore, achieving an efficient model for better investigation of the relationship between multiple combinations of lifestyle factors and mental health conditions.
Evaluation	How effective and accurate is the developed model in predicting mental health conditions based on lifestyle factors?	To evaluate the accuracy and effectiveness of the outcome of the relationship between multiple combinations of lifestyle factors with mental health conditions of adults in R Studio. Moreover, achieving visualizable models, to ensure a clear understanding of the relationships provided.

Table 3.1 Four Main Objectives in this Research Paper

IV. Literature Review

Mental health is an important component of overall well-being and quality of life, affecting emotional and psychological health. Based on the National Center for Health Statistics, about one in five adults (19.2%) in the United States received some form of mental health treatment in the year of 2019. Among them, 15.8% took medication and 9.5% attended counselling or therapy sessions with a mental health professional [2]. The figure of 19.2% represents a substantial number of adults receiving mental health treatment indicates that mental health issues are a widespread and serious concern. However, as not all individuals experiencing mental health issues have access to professional help, the actual prevalence of mental health issues is likely higher than reported. This highlights the importance of exploring broader determinants of mental health over clinical treatment data.

Apart from individual health outcomes, mental health has important social and economic implications. According to an article published in Public Health Reviews, poor mental health is related with social disconnection and social inequalities whereas good mental health supports human, social and economic development [3]. It mentions that people with poor mental health often experience reduced work performance and increased reliance on healthcare services. With these being said, this highlights the importance of understanding the factors that affect mental health and implementing effective strategies to promote well-being among individuals and communities.

In recent years, mental health issues are also strongly linked to severe and life-threatening outcomes, particularly suicide. It has emerged as a critical social problem linked to poor mental health. The World Health Organization (WHO) has revealed that more than 720,000 people die by suicide each year in worldwide [4]. Recent studies have shown that individuals with a current or past mental health issues are at the higher risk of suicide compared to those without such a diagnosis [5]. This demonstrates that mental health issues can contribute to serious and life-threatening effects if not properly addressed. Therefore, the need to identify modifiable factors that may help improve psychological well-being and reduce associated risks must be strengthened.

Lifestyle habits have been increasingly examined as potential determinants of mental health because they represent behaviours that individuals can modify. Lifestyle habits typically include physical activity such as exercise,

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

sleep duration, work-related behaviours, screen time and more. As such, public health research suggests that these behaviours may influence emotional regulation and stress level [4]. Thus, the seriousness of mental health issues must be emphasized by society.

Furthermore, there are several studies that have reported the positive associations between the specific lifestyle habits and mental well-being. For instance, physical activities have been linked to improved mood, reduced stress and higher life satisfaction across different age groups [6]. At the same time, a better emotional stability is associated with an adequate sleep duration while insufficient sleep has been linked to increased psychological distress. According to the study in Public Health Reviews, social interactions have also been widely recognised as a protective factor for mental health. This is due to it provides emotional support and contributes to higher levels of happiness and life satisfaction [7].

However, there is evidence regarding the relationship between lifestyle habits and mental health outcomes is not always consistent. Some studies have reported no statistical associations between related lifestyle measures and happiness. For example, an article published in the National Library of Medicine found out that there is no statistically significant relationship between dietary behaviour, diet quality and overall lifestyle scores to happiness levels after controlling for relevant factors [8]. Hence, this finding suggests that lifestyle habits may not always predict happiness in a statistically detectable way. It might depend on population characteristics or measurement methods.

In short, many studies focus on individual lifestyle factors rather than examining their combined effects. Moreover, the results of the studies show most of the lifestyle habits are associated with mental health outcomes such as depression, happiness or anxiety. This highlights the research gap in understanding whether multiple lifestyle habits are statistically associated with happiness outcomes among adults when considered the factors together. Addressing this gap provides the basis for the current studies which examines the collective relationship between lifestyle habits including exercise level, diet type, sleep hours, work hours per week, screen time per day and social interaction score with mental health outcomes measured using happiness scores.

V. DATASET SELECTION & JUSTIFICATION

A. Dataset Selection

Table 5.1 shows the detailed information on the selected dataset.

Component	Details
Dataset Name	Mental Health and Lifestyle Habits (2019-2024)
Publisher Name	Atharva Soundankar
Source of the Dataset	Kaggle
Dataset Link	https://www.kaggle.com/datasets/atharvasoundankar/mental-health-and-lifestyle-habits-2019-2024
Dataset Description	The Mental Health and Lifestyle Habits Datasets (2019-2024) are an extensive collection of data designed to understand the methods in which different lifestyle habit factors collectively impact mental health. The datasets collected all the respondents' basic demographic information, such as country of living, age and gender, with the essential aspects of the lifestyle factor, such as exercise routines, dietary habits, sleep patterns, stress levels, and social interactions.

Table 5.1 Detailed Information on the Selected Dataset

B. Key Features

In the selected dataset above, it consists of 3000 rows of data with 12 columns of features. Table 5.2 shows the detailed information on the 12 key features.

Feature Name	Data Type	Column Description
Country	Character	The country name of respondents' residents.
Age	Integer	The age of the respondent between when the data was being collected.
Gender	Categorical	The sex of the respondent, either male or female.
Exercise Level	Categorical	The frequency of respondents doing physical exercise, and categorized as Low, Moderate, or High.
Diet Type	Categorical	The dietary preference or pattern of respondents, either Balanced, Junk Food, Vegetarian, Vegan, or Keto.
Sleep Hours	Number	The average number of hours the respondent sleeps per night.
Stress Level	Categorical	The respondent self-reported their stress level, categorized as Low, Moderate, or High.
Mental Health Condition	Categorical	The mental health condition of the respondent, categorized as None, Anxiety, Depression, PTSD, and Bipolar.
Work Hours per Week	Integer	The average number of hours the respondents work per week.
Screen Time per Day (Hours)	Number	The average daily screen time in hours, including mobile phone, computer, and television use.
Social Interaction Score	Number	The score of the respondent's level of social engagement and connectedness.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

Happiness Score	Number	The respondent self-reported a happiness score based on their experience and feeling.

Table 5.2 Detailed Information on the 12 Key Features

C. Alignment with Objectives

The dataset is suitable to solve our problem statement stated above, as it contained all the essential lifestyle habit factors such as exercise level, diet type, sleep hours, work hours per week, screen time per day, and social interaction score, along with the dependent variable which is happiness score, identified as inversely proportional to the mental health of the adults. These features and variables directly support the objectives of this research study by allowing exploratory data analysis (EDA), statistical modeling, and predictive modeling to investigate and predict the relationships between lifestyle habit factors and mental health outcomes. In addition, the dataset was well-formatted to benefit for efficient data preprocessing, visualization, and evaluation in R Studio to achieve accurate, reliable, and interpretable results aligned with the research objectives and goals.

D. Preprocessing Needs / Challenges

In the dataset, there might be some minor error or missing data that needed to be completed to make the dataset more reliable and accurate.

1. **Check missing values**. This is to make sure there is zero missing value in the dataset, by validating the dataset and imputing the missing value if required.

2. **Data type and encoding needs**. Before applying the data to the predicting model, some character variables, such as Gender, Exercise Level, Diet Type, Stress Level, and Mental Health Condition, we must convert them into factors, and then perform the encoding process to encode them into numerical representation.

3. **Potential Outliers**. There might be some columns containing outliers which could distort the model performance. Therefore, boxplots or IQR method could be used to detect for any outlier in the dataset.

4. **Normalization and Standardization**. Continuous features such as Age, Sleep Hours, Works Hours per Week, Screen Time per Day, Social Interaction Score, and Happiness Score, vary in scale. Thus, to ensure all features contribute equally in the model, applying min-max scaling or StandardScaler is necessary.

5. **Feature Engineering**. Creating a new column that is meaningful and improves interpretability by categorizing the column's data to make it easy for understanding and comparing groups in descriptive or inferential statistics.

VI. EXPLORATORY DATA ANALYSIS (EDA)

Exploratory Data Analysis (EDA) is the process of observing, analyzing and investigating the raw data to outline their principal characteristics [9]. The EDA must be performed before the data preprocessing. However, to observe the EDA we need to load or import the dataset into the R and check the data structure of each variable inside the dataset beforehand, which is shown in the Section VII. Data Preprocessing part, where conducted in step 1 and 2.

Performing EDA was crucial for us, as a data analytics or statistician, as it can help us to understand the number of variables, the data type of each variable, and the distribution of each data [10]. Apart from that, to make our statistical result as accurate as possible, it can be performed by identifying the unusual data points, also known as outlier, in specific numerical variables using EDA. In addition, identifying invisible relationships and patterns between various data points could be beneficial when building the model. Moreover, getting insights from the EDA could also aid us to select the most appropriate and suitable techniques for modelling and modifying them for better outcomes.

A. Scatter Plots - Relationship between Two Numerical Variables

```
# EDA - Scatter plots
num_cols <- names(mental_health)[sapply(mental_health, is.numeric)]
num_cols <- setdiff(num_cols, "Happiness.Score")  # exclude target

for (col in num_cols) {
  print(
    ggplot(mental_health, aes_string(x = col, y = "Happiness.Score")) +
      geom_point(alpha = 0.6, color = "darkgreen") +
      geom_smooth(method = "lm", color = "red", se = FALSE, linewidth = 1) +
      theme_minimal() +
      ggtitle(paste("Scatter Plot:", col, "vs Happiness.Score"))
  )
}
```

Figure 6.1 Code for plotting scatter plots

Figure 6.1 shows the code for plotting a scatter plot for all the numeric independent variables, which included age, sleep hours, work hours per week, screen time per day (hours) and social interaction score, to show the relationship with the dependent variable, happiness score. Each variable that is numerical will be stored inside the 'num_cols' character vector as well as remove the happiness score variable from the 'num_cols' character vector since only the independent variable would be required to store inside the 'num_cols' character vector. Next will be iterated through each numeric independent variable as x-axis of the scatter plot diagram, with the fixed y-axis of happiness score. Finally the scatter plot will be printed out for each independent numeric variable against the dependent variable (happiness score).

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

Figure 6.2 Scatter plot for age vs happiness score

Figure 6.2 showing the scatter plot for the independent numeric variable (age) against the dependent variable (happiness score). The output for this plot was obviously no upward or downward pattern between these two variables, therefore it shows a zero correlation which means there is no relationship between these two variables.

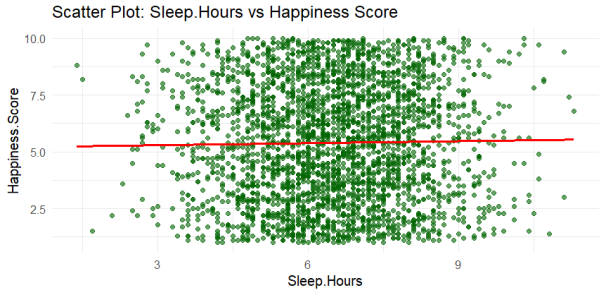

Figure 6.3 Scatter plot for sleep hours vs happiness score

Figure 6.3 showing the scatter plot for the independent numeric variable (sleep hours) against the dependent variable (happiness score). The output for this plot was obviously no upward or downward pattern between these two variables, therefore it shows a zero correlation which means there is no relationship between these two variables.

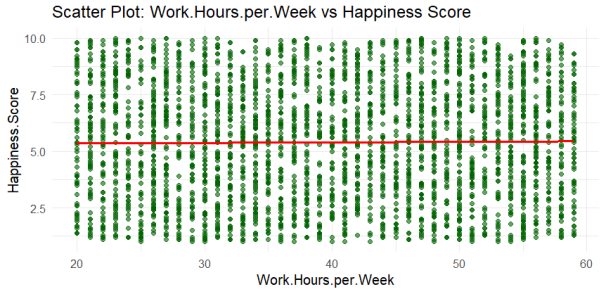

Figure 6.4 Scatter plot for work hours per week vs happiness score

Figure 6.4 showing the scatter plot for the independent numeric variable (work hours per week) against the dependent variable (happiness score). The output for this plot was obviously no upward or downward pattern between these two variables, therefore it shows a zero correlation which means there is no relationship between these two variables.

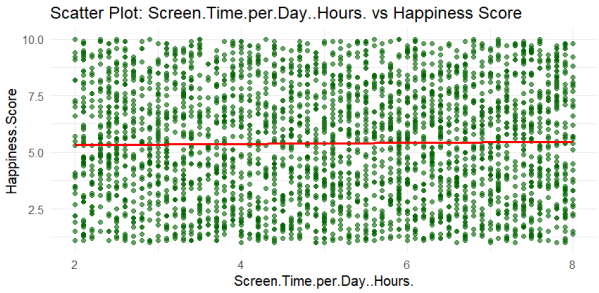

Figure 6.5 Scatter plot for screen time per day (hours) vs happiness score

Figure 6.5 showing the scatter plot for the independent numeric variable (screen time per day (hours)) against the dependent variable (happiness score). The output for this plot was obviously no upward or downward pattern between these two variables, therefore it shows a zero correlation which means there is no relationship between these two variables.

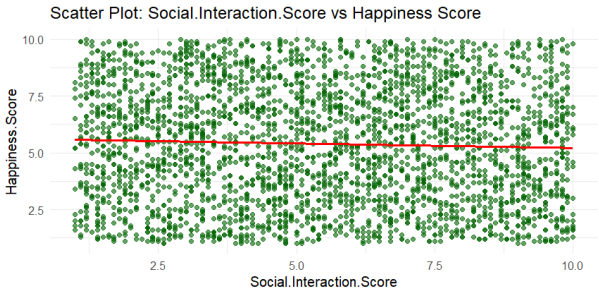

Figure 6.6 Scatter plot for social interaction score vs happiness score

Figure 6.6 showing the scatter plot for the independent numeric variable (social interaction score) against the dependent variable (happiness score). The output for this plot was obviously no upward or downward pattern between these two variables, therefore it shows a zero correlation which means there is no relationship between these two variables.

B. Bar Charts - Categorical Comparisons

```
# EDA - Bar charts
cat_cols <- names(mental_health)[sapply(mental_health, is.character)]

for (col in cat_cols) {
  print(
    ggplot(mental_health, aes_string(x = col)) +
      geom_bar(fill = "lightblue", color = "black") +
      ylab("Number of People") +
      theme_minimal() +
      ggtitle(paste("Bar Chart:", col, "Distribution")) +
      theme(axis.text.x = element_text(angle = 45, hjust = 1))
  )
}
```

Figure 6.7 Code for plotting bar charts

Figure 6.7 shows the code for plotting bar charts for each categorical variable against the number of people. The categorical variable in the dataset consists of country, gender, exercise level, diet type, stress level and mental health conditions. The dependent variable which is the number of people can be also known as the number of data points in a specified feature. Each variable that is categorical will be stored inside the 'cat_cols' character vector. Next will be iterated through each categorical variable as x-axis

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

of the bar chart, with the fixed y-axis of number of people. Finally the bar chart will be printed out for each categorical variable against the dependent variable (number of people).

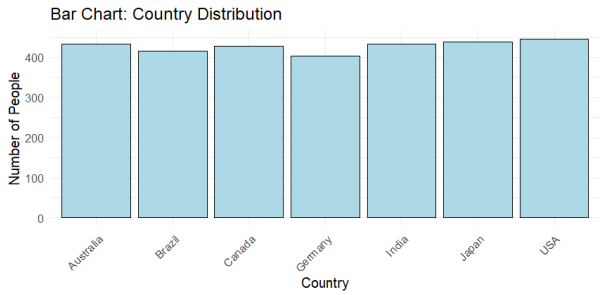

Figure 6.8 Bar chart for country vs number of people

Figure 6.8 shows the bar chart between the country and the number of people. By observing the output, the USA has the highest number of people, whereas Germany has the lowest number of people. The remaining countries have approximately a similar number of people living in the respective country.

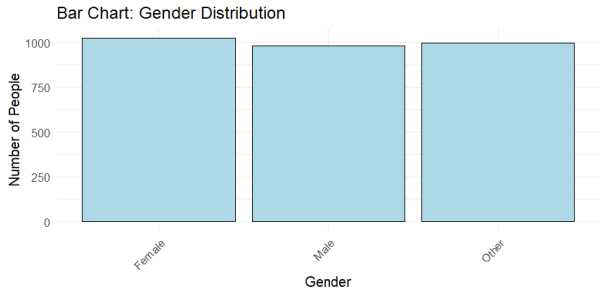

Figure 6.9 Bar chart for gender vs number of people

Figure 6.9 shows the bar chart between the gender and the number of people. By observing the output, females have the highest number of people among this dataset, whereas male have the lowest number of people among this dataset. There were around 1000 people who perhaps choose to not share their gender information under the 'Other' category.

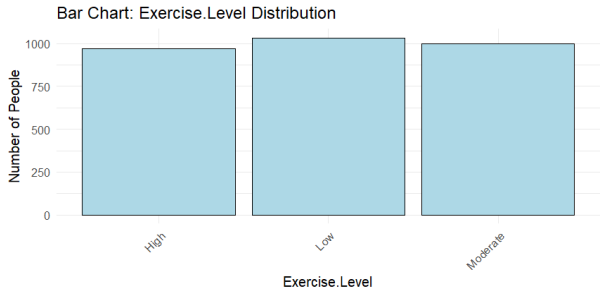

Figure 6.10 Bar chart for exercise level vs number of people

Figure 6.10 shows the bar chart between the exercise level and the number of people. By observing the output, the number of people who had a low exercise level was higher than the number of people who were having a high exercise level. The number of people who were having moderate exercise level was between the number of people who were having low exercise level and who were having high exercise level.

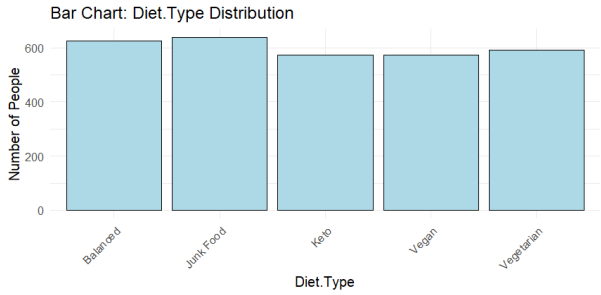

Figure 6.11 Bar chart for diet type vs number of people

Figure 6.11 shows the bar chart between the diet type and the number of people. By observing the output, people who took junk food was having the highest number of people compared with the others diet type, followed by balanced diet type, vegetarian, and keto and vegan, which was having the same number of people, with lowest number of people compared with others diet type.

Figure 6.12 Bar chart for stress level vs number of people

Figure 6.12 shows the bar chart between the stress level and the number of people. By observing the output, they were visually having the same number of people across different stress levels. However, when we take a detailed view, we can observe that people who have low stress level has the highest number of people, followed by people who having high stress level, and people who having moderate stress level.

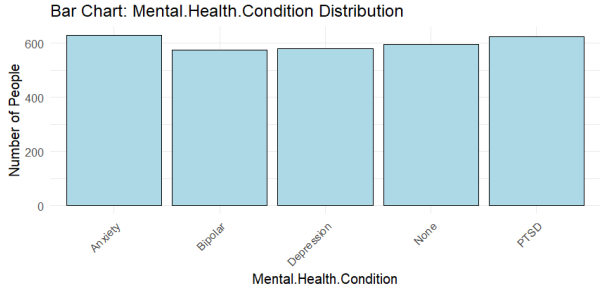

Figure 6.13 Bar chart for mental health condition vs number of people

Figure 6.13 shows the bar chart between the mental health condition and the number of people. By observing the output, people who were under anxiety condition were having the highest number of people, followed by under PTSD condition, under good mental health condition (None), under depression condition, and under bipolar

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

condition.

C. *Histograms - Distribution of a Single Numerical Variable*

```
# EDA - Histograms
num_cols <- names(mental_health)[sapply(mental_health, is.numeric)]

for (col in num_cols) {
  print(
    ggplot(mental_health, aes_string(x = col)) +
      geom_histogram(
        aes(y = after_stat(density)),
        bins = 30,
        fill = "skyblue",
        color = "black"
      ) +
      geom_density(color = "red", linewidth = 1) +
      ylab("Number of People") +
      theme_minimal() +
      ggtitle(paste("Histogram of", col))
  )
}
```

Figure 6.14 Code for plotting histograms

Figure 6.14 shows the code for plotting histograms for each single numerical variable against the number of people. The single numerical variable in the dataset consists of age, sleep hours, work hours per week, screen time per day (hours), and social interaction score. The dependent variable which is the number of people can be also known as the number of data points in a specified feature. Each variable that is a single numerical variable will be stored inside the 'num_cols' character vector. Next will be iterated through each single numerical variable as x-axis of the histogram, with the fixed y-axis of number of people. Finally the histogram will be printed out for each single numeric variable against the dependent variable (number of people).

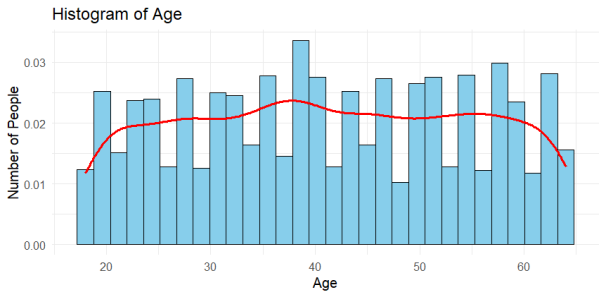

Figure 6.15 Histogram of age vs number of people

Figure 6.15 shows the histogram between the age and the number of people. By observing the output, the density curve that is in red colour does not show a clear distribution curve, which means that the age variable does not have a consistent and stable underlying distribution pattern in the data.

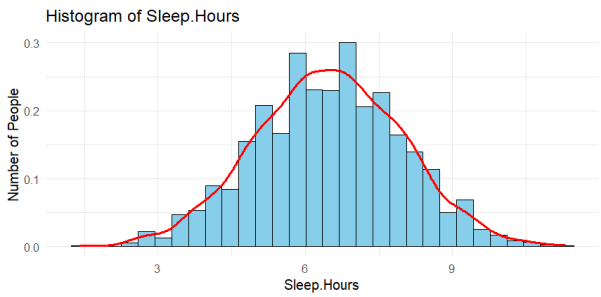

Figure 6.16 Histogram of sleep hours vs number of people

Figure 6.16 shows the histogram between the sleep hours and the number of people. By observing the output, the density curve that is in red colour shows a clear distribution curve, which means that the sleep hours variable has a consistent and stable underlying distribution pattern in the data.

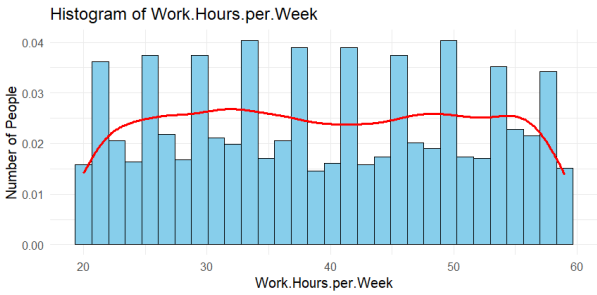

Figure5.17 Histogram of work hours per week vs number of people

Figure 6.17 shows the histogram between the work hours per week and the number of people. By observing the output, the density curve that is in red colour does not show a clear distribution curve, which means that the work hours per week variable does not have a consistent and stable underlying distribution pattern in the data.

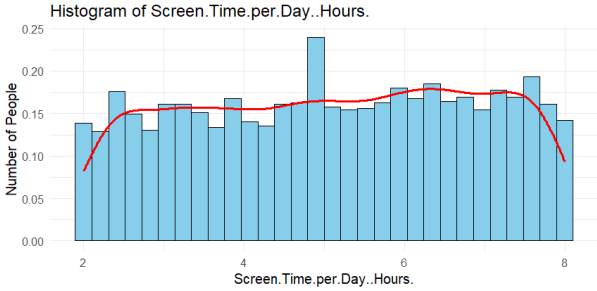

Figure 6.18 Histogram of screen time per day (hours) vs number of people

Figure 6.18 shows the histogram between the screen time per day (hours) and the number of people. By observing the output, the density curve that is in red colour does not show a clear distribution curve, which means that the screen time per day (hours) variable does not have a consistent and stable underlying distribution pattern in the data.

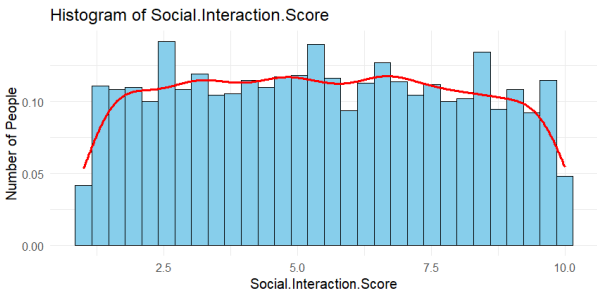

Figure 6.19 Histogram of social interaction score vs number of people

Figure 6.19 shows the histogram between the social interaction score and the number of people. By observing the output, the density curve that is in red colour does not show a clear distribution curve, which means that the social

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

interaction score variable does not have a consistent and stable underlying distribution pattern in the data.

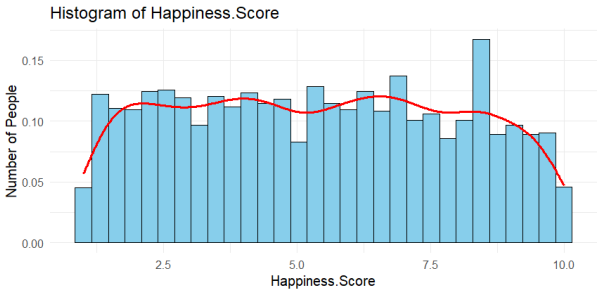

Figure 6.20 Histogram of happiness score vs number of people

Figure 6.20 shows the histogram between the happiness score and the number of people. By observing the output, the density curve that is in red colour does not show a clear distribution curve, which means that the happiness score variable does not have a consistent and stable underlying distribution pattern in the data.

D. Correlation Heatmap - Relationships between All Numerical Variables

```
# EDA - Correlation heatmaps
num_cols <- names(mental_health)[sapply(mental_health, is.numeric)]
cor_matrix <- cor(mental_health[, num_cols], use = "complete.obs")

ggcorrplot(cor_matrix,
           lab = TRUE,
           hc.order = TRUE,
           title = "Correlation Heatmap of Numeric Variables",
           ggtheme = theme_minimal())
```

Figure 6.21 Code for plotting correlation heatmap

Figure 6.21 shows the code for plotting correlation heatmap for all numerical variables. The numerical variable in the dataset consists of age, sleep hours, work hours per week, screen time per day (hours), social interaction score, and happiness score. Each variable that is a numerical variable will be stored inside the 'num_cols' character vector. Next, selects from the dataset that is numeric columns only, and uses the selected variables's rows that are not any missing values, to compute the pairwise correlation coefficients and store the calculated results in the 'cor_matrix' variable. Lastly, use the calculated correlation matrix to plot the correlation heatmap.

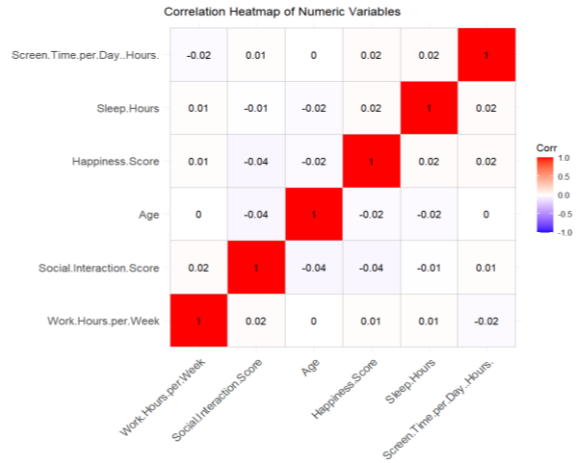

Figure 6.22 correlation heatmap between all numeric variables

Figure 6.22 shows the output of the correlation heatmap between all the numerical variables to identify how each numerical variable is closely related with other different numerical variables. By observing the correlation heatmap, there are some cells of value 1 with dark red colour appear along the diagonal of the matrix which have stronger correlations, because it is just comparing the relationship between one variable with the same variable, for example age variable has a correlation of 1 with itself. Therefore, the diagonal line of the red cells was represented as self-correlation. In terms of comparing the relationship between different variables, most of them have no meaningful linear relationship between the variables since the value is near to or even exactly 0 which means the variables do not move together in a linear way (6.3). In the technical term, the positive correlation value was having the extremely weak positive relationship, in contrast the negative correlation value was having the extremely weak negative relationship, but since both are closing to the correlation value of 0 so there is no any meaningful relationship between different variables.

VII. DATA PREPROCESSING

This section shows all the necessary steps to perform data preprocessing in R Studio.

A. Step 1: Load the library and dataset

```
# Load libraries
library(caret)
library(dplyr)

# Load dataset
mental_health <- read.csv("C:/Users/User/OneDrive/Statistical Inference and Modeling/Mental_Health_Lifestyle_Dataset.csv")
```

Figure 7.1 Load library and dataset (code)

The dataset is downloaded from Kaggle, named Mental Health and Lifestyle Habits (2019-2024). It is loaded in R using the read.csv() function.

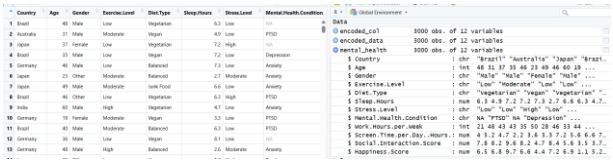

Figure 7.2 Dataframe of mental_health (result)

Figure 7.2 shows the output of the dataset, indicating that the dataset has been successfully loaded in R.

B. Step 2: Check Data Structure

```
# Check structure
str(mental_health)
```

Figure 7.3 Check Structure (code)

```
> str(mental_health)
'data.frame':   3000 obs. of  12 variables:
 $ Country                  : chr  "Brazil" "Australia" "Japan" "Brazil" ...
 $ Age                      : int  48 31 37 35 46 23 49 46 60 19 ...
 $ Gender                   : chr  "Male" "Male" "Female" "Male" ...
 $ Exercise.Level           : chr  "Low" "Moderate" "Low" "Low" ...
 $ Diet.Type                : chr  "Vegetarian" "Vegan" "Vegetarian" "Vegan" ...
 $ Sleep.Hours              : num  6.3 4.9 7.2 7.2 7.3 2.7 6.6 6.3 4.7 3.3 ...
 $ Stress.Level             : chr  "Low" "Low" "High" "Low" ...
 $ Mental.Health.Condition  : chr  NA "PTSD" NA "Depression" ...
 $ Work.Hours.per.Week      : int  21 48 43 43 35 50 28 46 33 44 ...
 $ Screen.Time.per.Day..Hours.: num  4 5.2 4.7 2.2 3.6 3.3 7.2 5.6 6.6 7.7 ...
 $ Social.Interaction.Score : num  7.8 8.2 9.6 8.2 4.7 8.4 5.6 3.5 3.7 3 ...
 $ Happiness.Score          : num  6.5 6.8 9.7 6.6 4.4 7.2 6.9 1.1 5.2 7.7 ...
```

Figure 7.4 Structure of each variable (result)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

The str() function is used to understand the structure of the mental_health data. The data types are confirmed to be suitable, and no conversion is needed.

C. Step 3: Handling Missing Values, Duplicate Rows, Extra Space, Bad Characters, and Removing NA Rows

```
# Check missing values
sum(is.na(mental_health))

# Handle duplicate rows
sum(duplicated(mental_health))

# Handle extra spaces in character columns
mental_health <- mental_health %>%
  mutate(across(where(is.character), trimws))

# Replace bad characters with NA
bad_chars <- c("", " ", "N/A", "NA", "n/a", "na", "?", "-", ".", "NULL", "Missing", "No data")
mental_health <- mental_health %>%
  mutate(across(where(is.character), ~ ifelse(. %in% bad_chars, NA, .)))
```

Figure 7.5 Handling Missing Values, Duplicate Rows, Extra Space, and Bad Characters (code)

The sum 'is.na()' function is used to check missing values in the dataset, while the sum(duplicated) function is applied to identify whether any duplicate rows exist. Extra spaces in character columns and bad characters are handled to ensure data consistency and accuracy before further analysis.

```
> # Check missing values
> sum(is.na(mental_health))
[1] 0
>
> # Handle duplicate rows
> sum(duplicated(mental_health))
[1] 0
```

Figure 7.6 Result of missing and duplicate rows (result)

Figure 7.6 shows that both functions return 0, indicating there are no missing values and no duplicated rows in the dataset. Hence, imputation and removal of duplicates are not performed.

```
# Remove rows with NA
mental_health <- na.omit(mental_health)
```

Figure 7.7 Remove NA rows (code)

All rows with missing or undefined values were removed from the dataset to ensure complete data for analysis and modeling.

D. Step 4: Encoding Categorical Variable

```
# Encoding Categorical Variable
encoded_data <- mental_health %>%
  mutate(across(where(is.character), ~ as.numeric(factor(.))))
```

Figure 7.8 Encoding Categorical Variable (code)

All categorical variables must be converted into a numeric format to create boxplots for each feature.

```
> head(encoded_data)
  Country Age Gender Exercise.Level Diet.Type Sleep.Hours Stress.Level
1       2  48      2              2         5         6.3            2
2       1  31      2              3         4         4.9            2
3       6  37      1              2         5         7.2            1
4       2  35      2              2         4         7.2            2
5       4  46      2              2         1         7.3            2
6       6  23      3              3         1         2.7            3
  Mental.Health.Condition Work.Hours.per.Week Screen.Time.per.Day..Hours.
1                      NA                  21                         4.0
2                       4                  48                         5.2
3                      NA                  43                         4.7
4                       3                  43                         2.2
5                       1                  35                         3.6
6                       1                  50                         3.3
  Social.Interaction.Score Happiness.Score
1                      7.8              6.5
2                      8.2              6.8
3                      9.6              9.7
4                      8.2              6.6
5                      4.7              4.4
6                      8.4              7.2
>
```

Figure 7.9 Show an overview of encoded_data (result)

Figure 7.9 shows categorical data converted into numeric codes.

E. Step 5: Handle Outlier

```
# Create boxplots
boxplot(encoded_data,
        main = "Boxplots of Mental Health and Lifestyle Variables",
        col = "orchid",
        las = 2,
        cex.axis = 0.8)
```

Figure 7.10 Create boxplot (code)

A boxplot is created to identify outliers within a dataset by visualizing the distribution of numeric data to detect extreme values.

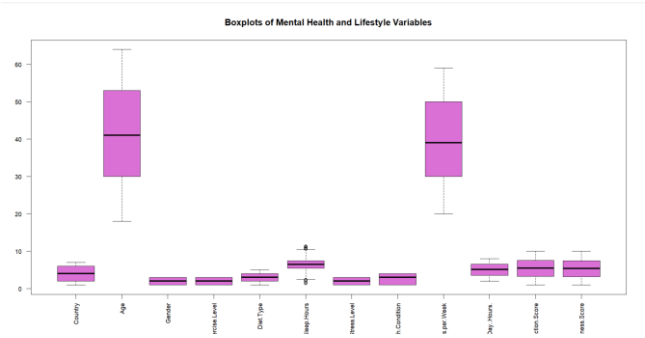

Figure 7.11 Boxplot showing outlier for Sleep Hours (result)

Figure 7.11 shows extreme values for Sleep Hours, indicating the presence of outliers. Therefore, only extreme values for Sleep Hours need to be handled to ensure the dataset remains reliable and does not bias the analysis.

```
# Handle outliers in Sleep.Hours using IQR method
Q1 <- quantile(mental_health$Sleep.Hours, 0.25, na.rm = TRUE)
Q3 <- quantile(mental_health$Sleep.Hours, 0.75, na.rm = TRUE)
IQR_val <- Q3 - Q1

lower_bound <- Q1 - 1.5 * IQR_val
upper_bound <- Q3 + 1.5 * IQR_val

# Cap outliers at the lower and upper bounds
mental_health$Sleep.Hours <- ifelse(mental_health$Sleep.Hours < lower_bound, lower_bound,
                                    ifelse(mental_health$Sleep.Hours > upper_bound, upper_bound,
                                           mental_health$Sleep.Hours))

# Boxplot after handling Sleep.Hours outliers
boxplot(mental_health$Sleep.Hours,
        main = "Boxplot of Sleep Hours After Outlier Handling",
        col = "lightgreen",
        las = 2,
        cex.axis = 0.8)
```

Figure 7.12 Handling outlier (code)

The extreme value of Sleep.Hours is handled using the Interquartile Range (IQR) capping method. Extreme values were capped at the lower and upper bounds.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

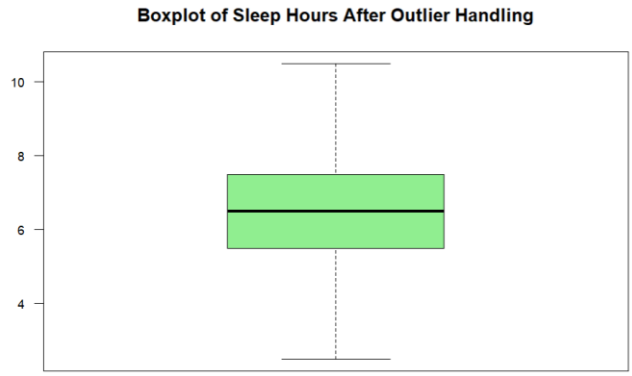

Figure 7.13 Boxplot of Sleep Hours after handling outlier (result)

Figure 7.13 shows the updated boxplot of Sleep Hours. The outliers have been successfully removed. It results in a more balanced distribution. This improves the dataset's suitability for statistical modeling by preventing extreme values from skewing the results.

F. Step 6: Feature Scaling

Scaling is not required for this dataset because it only involves times, hours and a scale up to 10 only, they do not range significantly.

G. Step 7: Feature Engineering

```
# Keep only relevant columns
mental_health<- mental_health[, c("Exercise.Level","Diet.Type","Sleep.Hours", "Mental.Health.Condition",
                                  "Work.Hours.per.Week","Screen.Time.per.Day..Hours.",
                                  "Social.Interaction.Score", "Happiness.Score")]

#Feature Engineering
mental_health$Happiness.Level <- as.factor(ifelse(mental_health$Happiness.Score
                                                  > median(mental_health$Happiness.Score), "High", "Low"))
```

Figure 7.14 Feature Engineering (code)

Other irrelevant columns were removed. A new categorical feature, Happiness.Level was created using ifelse() to label individuals with scores above the median as "High" and others as "Low." The result is converted to a factor for classification.

Happiness.Score	Happiness.Level
6.5	High
6.8	High
9.7	High
6.6	High
4.4	Low
7.2	High
6.9	High
1.1	Low
5.2	Low
7.7	High
5.5	High
8.0	High
5.1	Low
9.8	High
3.6	Low
1.8	Low

Figure 7.15 Happiness.Level column in scaled_data (result)

Happiness-level column is added to the dataframe. The dataset now has a binary target variable for classification. This simplifies modeling and allows evaluation using metrics such as accuracy or confusion matrices, while at the same time maintaining meaningful distinctions between higher and lower happiness levels instead of predicting continuous values.

H. Step 8: Train-test Split

```
#Convert characters to factor
mental_health$Exercise.Level <- as.factor(mental_health$Exercise.Level)
mental_health$Diet.Type <- as.factor(mental_health$Diet.Type)
mental_health$Mental.Health.Condition <- as.factor(mental_health$Mental.Health.Condition)


# Train-test split
set.seed(123)
trainIndex <- createDataPartition(mental_health$Happiness.Score, p = 0.8, list = FALSE)
train_data <- mental_health[trainIndex, ]
test_data <- mental_health[-trainIndex, ]
```

Figure 7.16 Train-test split (code)

Categorical features were converted to factors to ensure they are treated correctly in their correct type. The dataset is split into training (80%) and testing (20%) sets using createDataPartition() with a fixed seed for reproducibility. Stratified sampling ensures that the distribution of the target variable is preserved in both sets.

Exercise.Level	Diet.Type	Sleep.Hours	Mental.Health.Condition	Work.Hours.per.Week	Screen.Time.per.Day..Hours.	Social.Interaction.Score	Happiness.Score	Happiness.Level
3 Low	Vegetarian	7.2	NA	43	4.7	9.6	9.7	High
4 Low	Vegan	7.2	Depression	43	2.2	8.2	6.6	High
5 Low	Balanced	7.3	Anxiety	35	3.6	4.7	4.4	Low
6 Moderate	Balanced	2.7	Anxiety	50	3.3	8.4	7.2	High
7 Moderate	Junk Food	6.6	Anxiety	28	7.2	5.6	6.9	High
9 High	Vegetarian	4.7	Anxiety	33	6.6	3.7	5.2	Low
10 Moderate	Vegan	3.3	PTSD	44	7.7	3.0	7.7	High
11 Moderate	Balanced	6.3	PTSD	41	5.9	2.3	5.5	High
12 Low	Vegan	8.1	NA	54	7.9	2.7	8.0	High
13 High	Balanced	2.6	Anxiety	52	3.9	1.6	5.1	Low
14 Low	Vegan	5.3	Bipolar	46	3.1	9.0	9.8	High
15 Moderate	Balanced	6.7	Anxiety	59	4.4	9.4	3.6	Low
16 Moderate	Vegan	8.2	Anxiety	45	4.8	6.3	3.8	Low
17 Low	Balanced	6.0	Anxiety	38	7.4	8.6	4.9	Low
18 Low	Vegan	9.4	Bipolar	58	5.4	2.7	5.9	High
19 High	Vegetarian	3.8	Bipolar	45	3.8	9.1	4.1	Low
22 High	Junk Food	3.0	PTSD	25	3.0	2.3	4.4	Low
24 Moderate	Vegetarian	6.1	PTSD	22	7.5	1.4	8.0	High
26 Moderate	Junk Food	5.9	Depression	32	5.5	1.5	4.4	Low

Figure 7.17 Train_data dataframe (result)

Exercise.Level	Diet.Type	Sleep.Hours	Mental.Health.Condition	Work.Hours.per.Week	Screen.Time.per.Day..Hours.	Social.Interaction.Score	Happiness.Score	Happiness.Level
1 Low	Vegetarian	6.3	NA	21	4.0	7.8	6.5	High
2 Moderate	Vegan	4.9	PTSD	48	5.2	8.2	6.8	High
8 Low	Vegetarian	6.3	PTSD	46	5.6	3.5	1.1	Low
20 High	Junk Food	3.5	PTSD	25	6.5	5.6	3.2	Low
21 Moderate	Vegan	8.0	Bipolar	22	5.0	1.6	7.0	High
23 High	Vegetarian	7.2	NA	47	5.5	3.8	6.7	High
25 Moderate	Balanced	5.1	Anxiety	32	7.1	3.8	7.8	High
31 Moderate	Vegan	5.6	Bipolar	50	4.2	3.2	4.1	Low
39 Moderate	Junk Food	6.1	Bipolar	50	5.7	6.5	9.4	High
45 Low	Keto	7.2	Bipolar	45	6.3	5.1	4.2	Low
46 High	Junk Food	7.8	Bipolar	45	7.6	6.0	4.8	Low
48 Low	Vegetarian	6.5	NA	34	2.3	3.1	5.7	High
55 Low	Junk Food	8.0	NA	45	3.3	2.4	2.4	Low
60 Moderate	Keto	3.9	PTSD	37	3.9	7.5	6.2	High
69 Low	Vegan	4.9	Anxiety	50	2.0	3.8	10.0	High
71 High	Balanced	5.9	NA	59	5.2	4.0	8.9	High
80 High	Junk Food	3.6	NA	39	4.2	7.8	1.3	Low
86 Moderate	Keto	7.3	Depression	23	3.3	8.2	7.5	High
87 Moderate	Keto	6.5	Anxiety	46	7.6	4.4	3.5	Low

Figure 7.18 Test_data dataframe (result)

Out of 3000 total records, 2401 are used for training and 599 for testing. The training set has one extra row because createDataPartition() preserves class proportions, so sometimes the counts are rounded. This ensures that the model is evaluated on unseen data to improve the reliability of the results.

VIII. HYPOTHESIS FORMULATION

Hypothesis 1: Exercise Level and Happiness Score (One-way ANOVA)
This set tests if the average happiness score differs across different exercise levels.
H0 (Null Hypothesis): There is no significant difference in the mean of happiness score across the three exercise levels (Low, Moderate, High).
Ha (Alternative Hypothesis): At least one exercise level has a significantly different mean happiness score compared to the others.

Hypothesis 2: Type of Eaters and Happiness Score (Two-Sample T-test)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

This set tests if different types of eaters influence adults' average happiness score.
H0 (Null Hypothesis): The mean of happiness score is the same for plant-based eaters (vegan and vegetarian) and non-plant-based eaters (balanced, junk food and keto).
Ha (Alternative Hypothesis): The mean of happiness score is significantly different between plant-based eaters (vegan and vegetarian) and non-plant-based eaters (balanced, junk food and keto).

Hypothesis 3: Social Interaction Score and Happiness Score (Pearson Correlation)
This set tests the linear relationship between an adult's social interaction score and happiness
score statistically.
H0 (Null Hypothesis): There is no significant linear correlation between social interaction score and happiness score.
Ha (Alternative Hypothesis): There is a significant linear correlation between social interaction score and happiness score.

Hypothesis 4: Work Hours per Week and Happiness Score (Pearson Correlation)
This set examines if work hours per week and happiness score have a statistically significant linear connection.
H_0 (Null Hypothesis): There is no linear relationship between work hours per week and happiness score.
Ha (Alternative Hypothesis): A significant linear relationship exists between work hours per week and happiness score.

Hypothesis 5: Screen Time Level and Happiness Score (One-way ANOVA)
This set examines whether the average happiness score remains the same across different screen time levels.
H0 (Null Hypothesis): The mean of happiness score does not significantly differ between three screen time levels (Low, Medium, High).
Ha (Alternative Hypothesis): At least one screen time level has a significantly different mean happiness score compared to the others.

Hypothesis 6: Sleep Hours and Happiness Score (Pearson Correlation)
This set tests the statistical significant linear relationship between sleep hours and happiness score.
H0 (Null Hypothesis): There is no significant linear correlation between sleep hours and happiness score.
Ha (Alternative Hypothesis): There is a significant linear correlation between sleep hours and happiness score.

IX. STATISTICAL TESTING

A. Statistical Test Selection

Three statistical tests were conducted for the six hypotheses. The one way ANOVA was used to test hypothesis 1 and 5, the Pearson correlation was applied to hypothesis 3, 4 and 6, and the two samples T-test was used for hypothesis 2.

1) One-way ANOVA test

The One-way ANOVA test is an analysis of variance used to compare the means of multiple groups. The screen time per day hours variable was divided into three screen time levels：low, medium, high, and the exercise level also consists of three groups: low, moderate, high. Both hypothesis 1 and 5 involve a categorical independent variable and a numeric dependent variable (happiness score). Due to this structure, the one way ANOVA is suitable for hypothesis 1 and 5 to determine whether the average happiness score is the same for all the different levels.

2) Pearson Correlation

Pearson correlation can efficiently measure the linear relationship between two numeric variables. Hypotheses 3, 4 and 6 involve one numeric independent variable (social interaction score, work hours per week and sleep hours) and one numeric dependent variable (happiness score). Therefore, Pearson correlation is suitable to test if a linear relationship exists between two numeric variables for these hypotheses.

3) Two Sample T-test

The Two Sample T-test is suitable to be used for comparing categorical variables with two different groups and numerical variables. For Hypothesis 2, it has one categorical independent variable which is diet type, with two different groups, which is plant-based eaters (vegan and vegetarian) and non-plant-based eaters (balanced, junk food and keto), comparing with one numerical dependent variable, which is the happiness score. Two Sample T-tests can efficiently perform testing to test if different diet types can affect the average happiness score, or does not affect the average happiness score.

B. Verification of Assumption Fulfilment

1) Evaluate the no multicollinearity assumption

```
> # assumption validation - no multicollinearity
> model <- lm(Happiness.Score ~ Exercise.Level + Diet.Type + Sleep.Hours + Work.Hours.per.Week +
+               Screen.Time.per.Day..Hours. + Social.Interaction.Score, data = mental_health)
> vif(model)
                              GVIF Df GVIF^(1/(2*Df))
Exercise.Level            1.006529  2        1.001628
Diet.Type                 1.008191  4        1.001020
Sleep.Hours               1.003016  1        1.001507
Work.Hours.per.Week       1.002068  1        1.001033
Screen.Time.per.Day..Hours. 1.003281  1        1.001639
Social.Interaction.Score  1.001302  1        1.000651
```

Figure 9.1 GVIF of each features

Multicollinearity happens when two or more independent variables are highly correlated with each other. This situation should be avoided in order to ensure an accurate and reliable result of further statistical tests and predictive modelling. A variance inflation factor (VIF) that indicates how strongly each predictor is correlated with each other was used to verify the fulfilment of the no multicollinearity assumption. Figure 9.1 shows that the vif() function automatically provides generalized VIFs as the model includes two categorical independent variables. The result shows that all independent variables have GVIF below 1.5, indicating that multicollinearity is not a concern [12]. Thus, this assumption was fulfilled.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

2) Diagnostic Plots

```
# Create linear model
happiness.lm <- lm(Happiness.Score ~ Exercise.Level + Diet.Type + Sleep.Hours + Work.Hours.per.Week +
                   Screen.Time.per.Day..Hours. + Social.Interaction.Score, data = train_data)

# Check visualization of assumption
par(mfrow = c(2,2))
plot(happiness.lm)
par(mfrow = c(1,1))
```

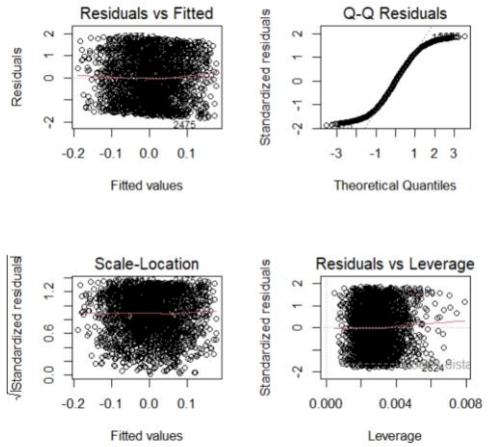

Figure 9.2: Diagnostic Plot

Based on Figure 9.2, a multiple linear model was created for checking the assumption fulfilment. Therefore, after fitting the model, the diagnostic plot will be created as shown in Figure 9.2, it contains residuals vs fitted plot, Q-Q plot, scale-location plot and residuals vs leverage plots. The purpose of these plots is to check on assumptions like linearity, normality and homoscedasticity, thus, it can determine the suitability of the model and data for further statistical testing and modeling.

The Residuals vs Fitted plot is used to examine the relationship between variables, in the diagram shown in Figure 9.2, it shows that the most residuals appears as a random scatter around the horizontal line at zero, and showing no visible curve, hence the **linearity** assumption was met.

The Q-Q Residuals plot is used to check if the residuals follows a normal distribution, in the diagram shown in Figure 9.2, although it has shown systematic deviation at both of the tails of the reference line, but as the dataset contains large data, it is roughly acceptable and as most of the residual points aligns with the reference line, this shows that the residuals follows a normal distribution approximately, hence the **normality** assumption was met.
Based on the Scale-Location plot, the residual variance points and the red line have roughly formed a horizontal band, it shows the constant or equal variance between variables, hence the **homoscedasticity** assumption is reasonably met.

Therefore, the Residuals vs Leverage plot has shown that most of the residual points have lower leverage but some of them are roughly near to the Cook's distance contours, showing that there might be variables that needed further inspection but does not affect statistical testing and modeling.

C. Perform Statistical Tests

1) Hypothesis 1: One way ANOVA

a) Perform One way ANOVA

```
> #Hypothesis 1: Exercise Level and Happiness Score (ANOVA)
>
> # Encode categorical variable (Exercise variable) to factor
> mental_health$Exercise.Level <- factor(mental_health$Exercise.Level)
>
> # one way anova
> one_way_anova_h1 <- aov(Happiness.Score ~ Exercise.Level, data = mental_health)
> summary(one_way_anova_h1)
                Df Sum Sq Mean Sq F value Pr(>F)
Exercise.Level    2     35  17.376   2.659 0.0702 .
Residuals      2997  19583   6.534
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Figure 9.3 One way ANOVA for Hypothesis 1

The categorical variable, exercise level was first encoded into a factor. The one-way ANOVA was performed using the aov() function, the result was summarized using the summary() function. The summary shows the degrees of freedom, sum of squares, mean square, F-statistic and p-value.

b) Tukey HSD post hoc test

```
> # Post-hoc test
> tukey.one.way.h1<-TukeyHSD(one_way_anova_h1)
> tukey.one.way.h1
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = Happiness.Score ~ Exercise.Level, data = mental_health)

$Exercise.Level
                     diff        lwr        upr     p adj
Low-High      -0.25586262 -0.5239220 0.01219676 0.0650878
Moderate-High -0.18709100 -0.4574159 0.08323389 0.2360899
Moderate-Low   0.06877162 -0.1972706 0.33481383 0.8167723
```

Figure 9.4 Tukey HSD post hoc test for Hypothesis 1

One way ANOVA only indicates that at least one group differs, thus, an additional Tukey HSD post hoc test was performed to examine pairwise differences. The post hoc test shows the 95% confidence intervals (diff - lwr), adjusted p-values, and the pairwise mean differences between different exercise levels.

2) Hypothesis 2: Two Sample T-test

a) Creating New Feature: Type of Eaters

To perform the Two Sample T-test to investigate if different diet types influence adult's average happiness score, the variable Diet Type should have two separate groups which is plant-based eaters (vegan and vegetarian) and non-plant-based eaters (balanced, junk food and keto), hence, to perform the test, the new feature type named "Type of Eaters" is created as shown in Figure 9.5, separating categories to plant-based and non-plant-based eaters.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

```
> # Create new feature based on Diet.Type variable
> mental_health$Type.of.Eaters <- ifelse(mental_health$Diet.Type %in% c("Vegan", "Vegetarian"),
+                                        "Plant-based",
+                                        "Non-plant-based")
>
> # Convert to factor
> mental_health$Type.of.Eaters <- as.factor(mental_health$Type.of.Eaters)
> head(mental_health)
   Country Age Gender Exercise.Level  Diet.Type Sleep.Hours Stress.Level Mental.Health.Condition Work.Hours.per.Week
1   Brazil  48   Male            Low Vegetarian         6.3          Low                    <NA>                  21
2 Australia  31   Male       Moderate      Vegan         4.9          Low                    PTSD                  48
3    Japan  37 Female            Low Vegetarian         7.2         High                    <NA>                  43
4   Brazil  35   Male            Low      Vegan         7.2          Low              Depression                  43
5  Germany  46   Male            Low   Balanced         7.3          Low                 Anxiety                  35
6    Japan  23  Other       Moderate   Balanced         2.7     Moderate                 Anxiety                  50
  Screen.Time.per.Day..Hours. Social.Interaction.Score Happiness.Score  Type.of.Eaters
1                         4.0                      7.8             6.5     Plant-based
2                         5.2                      8.2             6.8     Plant-based
3                         4.7                      9.6             9.7     Plant-based
4                         2.2                      8.2             6.6     Plant-based
5                         3.6                      4.7             4.4 Non-plant-based
6                         3.3                      8.4             7.2 Non-plant-based
>
```

Figure 9.5: New Feature Type of Eaters

b) Perform Two Sample T-test

```
> #Two Sample T-test
> t_test_result <- t.test(Happiness.Score ~ Type.of.Eaters,
+                         data = mental_health,
+                         var.equal = TRUE)
>
> t_test_result

        Two Sample t-test

data:  Happiness.Score by Type.of.Eaters
t = -1.4305, df = 2998, p-value = 0.1527
alternative hypothesis: true difference in means between group Non-plant-based and group Plant-based is not equal to 0
95 percent confidence interval:
 -0.32485917  0.05079666
sample estimates:
mean in group Non-plant-based     mean in group Plant-based
                     5.341853                      5.478884
```

Figure 9.6: Code of Performing Two Sample T-test for Hypothesis 2

Figure 9.6 above shows the Two Sample T-test performed for Hypothesis 2, and the outcome is as shown below:

1. **Mean**: The group Non-plant-based from the variable Type of Eaters have lower mean of happiness score which is 5.342, and the group Plant-based have higher mean of happiness score, which is 5.489.
2. **p-value**: The p-value is 0.1527.
3. **Confidence Interval**: The 95% confidence interval for both groups is ranged from -0.325 to 0.051.

Hypothesis 3: Pearson Correlation

c) Correlation between Social Interaction Score and Happiness Score

```
> #Hypothesis 3: Social Interaction Score and Happiness Score (Pearson Correlation)
>
> # visualizing data
> ggscatter(mental_health, x = "Social.Interaction.Score", y = "Happiness.Score",
+           add = "reg.line", conf.int = TRUE,
+           cor.coef = TRUE, cor.method = "pearson",
+           xlab = "Social.Interaction.Score", ylab = "Happiness.Score",alpha = 0.5, size = 1.5)
```

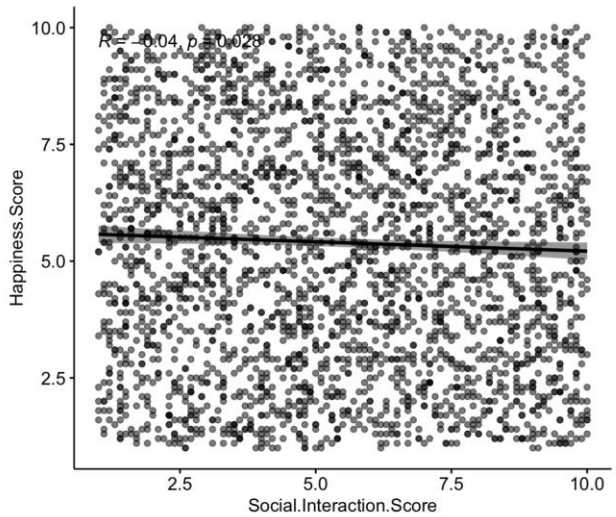

Figure 9.7 Scatter plot: correlation between social interaction score and happiness score

A scatter plot was created to visualize the relationship between social interaction score and happiness score. It shows the regression line with its 95% confidence interval and the Pearson correlation results to identify the strength of the linear relationship.

d) Pearson Correlation Test

```
> # Pearson Correlation
> Pearson_result_h3 <- cor.test(mental_health$Social.Interaction.Score, mental_health$Happiness.Score,
+                               method = "pearson")
> Pearson_result_h3

        Pearson's product-moment correlation

data:  mental_health$Social.Interaction.Score and mental_health$Happiness.Score
t = -2.2049, df = 2998, p-value = 0.02754
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.075912880 -0.004455603
sample estimates:
       cor
-0.04023569
```

Figure 9.8 Pearson Correlation for Hypothesis 3

Based on Figure 9.8, the Pearson correlation test for Hypothesis 3 to examine the linear relationship between social interaction score and happiness score was conducted. The result for the test is shown below:

1. **p-value**: The p-value is 0.02754.
2. **Confidence Interval**: The 95% confidence interval is ranged from -0.076 to -0.004.

3) Hypothesis 4: Pearson Correlation

a) Correlation between Work Hours per Week and Happiness Score

```
> #Hypothesis 4: Pearson (Work Hours per Week and Happiness Score)
> ggscatter(mental_health, x = "Work.Hours.per.Week", y = "Happiness.Score",
+           add = "reg.line", conf.int = TRUE,
+           cor.coef = TRUE, cor.method = "pearson",
+           xlab = "Work.Hours.per.Week", ylab = "Happiness.Score", alpha = 0.5, size = 1.5)
>
```

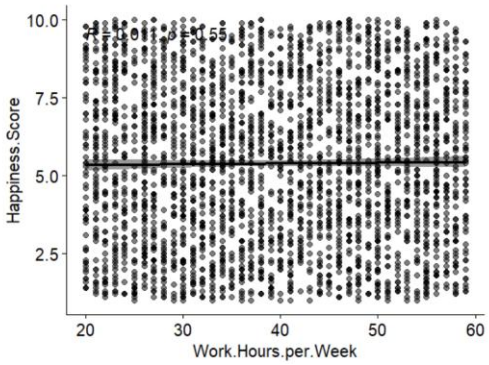

Figure 9.9 Scatter plot: correlation between work hours per week and happiness score

A scatter plot was created to visualize the relationship between work hours per week and happiness score based on Figure 9.9. It shows the regression line with its 95% confidence interval and the Pearson correlation results to identify the strength of the linear relationship.

b) Pearson Correlation Test

```
> cor_work <- cor.test(mental_health$Work.Hours.per.Week, mental_health$Happiness.Score, method = "pearson")
> cor_work

        Pearson's product-moment correlation

data:  mental_health$Work.Hours.per.Week and mental_health$Happiness.Score
t = 0.59342, df = 2998, p-value = 0.5529
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.02495886  0.04660575
sample estimates:
       cor
0.01083732
```

Figure 9.10: Code of Performing Pearson Correlation Test for Hypothesis 4

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

Based on Figure 9.10, the Pearson correlation test for Hypothesis 4, which is for investigating the relationship of Work Hours per Week and the happiness score was conducted, and the outcome is shown below:

1. **p-value**: The p-value is 0.5529.
2. **Confidence Interval**: The 95% confidence interval is ranged from -0.025 to 0.047.

4) Hypothesis 5: One way ANOVA

a) Create screen time level feature

```
> #Create ScreenTime.level feature
> mental_health$ScreenTime.level <- cut(mental_health$Screen.Time.per.Day..Hours.,
+                                       breaks = 3,
+                                       labels = c("Low", "Medium", "High"),
+                                       include.lowest = TRUE)
```

Figure 9.11 Create ScreenTime.level variable

```
> head(mental_health)
    Country Age Gender Exercise.Level  Diet.Type Sleep.Hours Stress.Level Mental.Health.Condition
1    Brazil  48   Male            Low Vegetarian         6.3          Low                    <NA>
2 Australia  31   Male       Moderate      Vegan         4.9          Low                    PTSD
3     Japan  37 Female            Low Vegetarian         7.2         High                    <NA>
4    Brazil  35   Male            Low      Vegan         7.2          Low              Depression
5   Germany  46   Male            Low   Balanced         7.3          Low                 Anxiety
6     Japan  23  Other       Moderate   Balanced         2.7     Moderate                 Anxiety
  Work.Hours.per.Week Screen.Time.per.Day..Hours. Social.Interaction.Score Happiness.Score ScreenTime.level
1                  21                         4.0                      7.8             6.5              Low
2                  48                         5.2                      8.2             6.8           Medium
3                  43                         4.7                      9.6             9.7           Medium
4                  43                         2.2                      8.2             6.6              Low
5                  35                         3.6                      4.7             4.4              Low
6                  50                         3.3                      8.4             7.2              Low
```

Figure 9.12 Data with newly created variable

ScreenTime.level was created by dividing the screen time per day hours of each observation into 3 equal portions, then assigning them to the Low, Medium and High screen time level (see Figure 9.11). The head() function was used to show the first few rows of data with the newly created variable. This new variable was created for hypothesis 4 as a categorical independent variable is required to perform the one way ANOVA test.

b) Perform One way ANOVA

```
> #Hypothesis 5: Screen Time Level and Happiness Score (One-way ANOVA)
>
> # one way anova
> one_way_anova_h5 <- aov(Happiness.Score ~ ScreenTime.level, data = mental_health)
> summary(one_way_anova_h5)
                  Df Sum Sq Mean Sq F value Pr(>F)
ScreenTime.level    2      9   4.528   0.692  0.501
Residuals        2997  19608   6.543
```

Figure 9.13 One way ANOVA for Hypothesis 5

The one-way ANOVA of hypothesis 5 was performed using the aov() function. The result was summarized using the summary function with the ANOVA object, providing the degrees of freedom, sum of squares, mean square, F-statistic and p-value for the effect of screen time level on happiness score.

c) Tukey HSD post hoc test

```
> # Post-hoc test
> tukey.one.way.h5<-TukeyHSD(one_way_anova_h5)
> tukey.one.way.h5
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = Happiness.Score ~ ScreenTime.level, data = mental_health)

$ScreenTime.level
                    diff        lwr       upr     p adj
Medium-Low  -0.02442752 -0.2957300 0.2468750 0.9757267
High-Low     0.10120034 -0.1658610 0.3682617 0.6475765
High-Medium  0.12562787 -0.1410808 0.3923365 0.5113382
```

Figure 9.14 Tukey HSD post hoc test for Hypothesis 5

As ANOVA only shows whether any difference exists among the groups, therefore, a Tukey HSD post hoc test was performed to examine pairwise differences. The post hoc test shows the pairwise ai4mean differences between the levels of screen time levels, the 95% confidence intervals (diff - lwr) and adjusted p-values.

5) Hypothesis 6: Pearson Correlation

a) Correlation between Sleep Hours and Happiness Score

```
> #Hypothesis 6: Pearson (Sleep Hours and Happiness Score)
> ggscatter(mental_health, x = "Sleep.Hours", y = "Happiness.Score",
+           add = "reg.line", conf.int = TRUE,
+           cor.coef = TRUE, cor.method = "pearson",
+           xlab = "Sleep.Hours", ylab = "Happiness.Score", alpha = 0.5, size = 1.5)
```

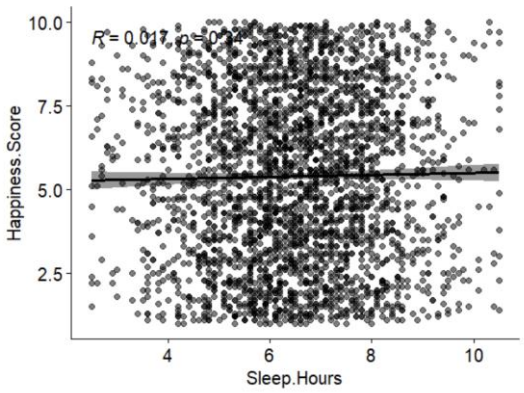

Figure 9.15 Scatter plot: correlation between sleep hours and happiness score

A scatter plot was created to visualize the relationship between sleep hours and happiness score based on Figure 9.15. It shows the regression line with its 95% confidence interval and the Pearson correlation results to identify the strength of the linear relationship.

b) Pearson Correlation Test

```
> cor_sleep <- cor.test(mental_health$Sleep.Hours, mental_health$Happiness.Score, method = "pearson")
> cor_sleep

        Pearson's product-moment correlation

data:  mental_health$Sleep.Hours and mental_health$Happiness.Score
t = 0.94862, df = 2998, p-value = 0.3429
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.01847548  0.05307607
sample estimates:
       cor
0.01732247
```

Figure 9.16: Code of Performing Pearson Correlation Test for Hypothesis 6

Based on Figure 9.16, the pearson correlation test for Hypothesis 6 was conducted which is for investigating the relationship between Sleep Hours and Happiness Score, and the outcome is shown below:

1. **p-value**: The p-value is 0.3429.
2. **Confidence Interval**: The 95% confidence interval is ranged from -0.018 to 0.053.

D. Interpretation of Results

1) One way ANOVA result - Hypothesis 1: Exercise Level and Happiness Score

Based on the result of a one-way ANOVA test shown in Figure 9.3, the F value was 2.659, and the p value was

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

0.0702, which is larger than the significance level 0.05. The result of Tukey HSD post hoc test shows that all pairwise comparisons have p-value > 0.05, and all the 95% confidence intervals (lwr - upr) were included zero. Both results of ANOVA and post hoc test shows that it fails to reject the null hypothesis, indicating that there is no significant difference in the mean of happiness score across the three exercise levels.

2) Two Sample T-test result - Hypothesis 2: Type of Eaters and Happiness Score

By analyzing the output of the Two Sample T-test for Hypothesis 2 as shown in Figure 9.6, the p-value is 0.1527, which is larger than the significance level 0.05. Therefore, the 95% confidence interval for both groups is ranged from -0.325 to 0.051, which includes 0 in the range, thus, it showed that it fails to reject the null hypothesis. This means that the mean of the happiness score is the same for plant-based eaters and non-plant-based eaters, showing no significant difference between both groups.

3) Pearson correlation result - Hypothesis 3: Social Interaction Score and Happiness Score

The scatter plot of social interaction score and happiness score (see Figure 9.7) shows that the data points were distributed randomly. The regression line is also nearly horizontal, which does not explicitly form either a positive or negative regression line. Moreover, based on the result of Pearson correlation test shown in Figure 9.8, the p-value (0.02754) was less than the significant level alpha (0.05). However, the correlation coefficient (R = -0.04) was close to zero, and even though the 95% confidence interval for the correlation range between -0.07591 and -0.004, which does not contain zero, it was actually close to zero. Therefore, the result shows that there is a statistically significant but very weak linear correlation between social interaction score and happiness score, indicating that null hypothesis 3 is rejected.

4) Pearson correlation result - Hypothesis 4: Work Hours per Week and Happiness Score

The scatter plot of work hours per week and happiness score (see Figure 9.9) shows that the data points were distributed randomly. Based on the analysis of the result of the Pearson Correlation test for Hypothesis 4, as shown in Figure 9.10, the p-value is 0.5529, which is larger than the significance level 0.05. Moreover, the 95% confidence interval is ranged from -0.025 to 0.047, with 0 included in the range, thus, it shows that it fails to reject the null hypothesis. This means that there is no linear relationship between work hours per week and happiness score.

5) One way ANOVA result - Hypothesis 5: Screen Time Level and Happiness Score

As the result of a one-way ANOVA test shown in Figure 9.12, the F value was 0.692, and the p value was 0.501, which is larger than the significance level 0.05. Furthermore, the result of Tukey HSD post hoc test shows that all pairwise comparisons have p-value > 0.05, and all the 95% confidence intervals (lwr - upr) were included zero. The results of ANOVA and post hoc test shows that it fails to reject the null hypothesis, showing that there is no significant difference in the mean of happiness score among the three screen time levels.

6) Pearson correlation result - Hypothesis 6: Sleep Hours and Happiness Score

The scatter plot of sleep hours and happiness score (see Figure 9.15) shows that the data points were distributed randomly. By analyzing the result of the Pearson Correlation test for Hypothesis 6, as shown in Figure 9.16, the p-value is 0.3429, which is larger than the significance level 0.05. Moreover, the 95% confidence interval is ranged from -0.018 to 0.053, and it includes 0 in the range, thus, it shows that it fails to reject the null hypothesis. This means that there is no linear relationship between sleep hours and happiness score.

X. REGRESSION ANALYSIS

A. *Model Selection*

Random Forest Regression (RFR) will be used because it is a flexible and robust model that handles non-linear relationships and captures feature interactions automatically [13].

Figure 10.2 confirms the lack of linear correlation. In all four cases, the data points are widely dispersed with no pattern and the regression lines are nearly horizontal. Therefore, RFR was more suitable to use in our case rather than the multiple linear regression (MLR).

```
#Test Linearity
# List numeric variable
numeric_vars <- c("Sleep.Hours", "Work.Hours.per.Week",
                  "Screen.Time.per.Day..Hours.", "Social.Interaction.Score")

# Scatterplots with linear fit
par(mfrow=c(2,3))
for (var in numeric_vars) {
  plot(train_data[[var]], train_data$Happiness.Score,
       main = paste("Happiness Score vs", var),
       xlab = var, ylab = "Happiness Score", col="blue", pch=16)
  abline(lm(Happiness.Score ~ train_data[[var]], data=train_data), col="red")
}
par(mfrow=c(1,1))
```

Figure 10.1 Code of Testing Linearity

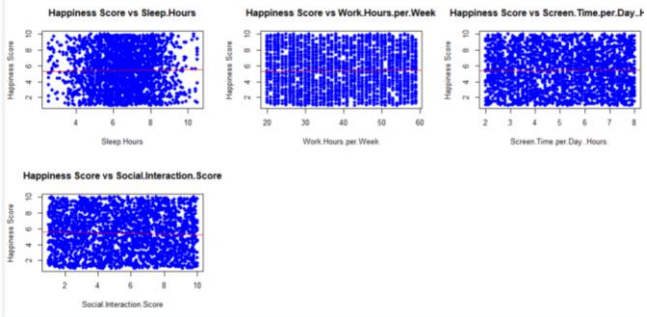

Figure 10.2 Result of the Linearity Test

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

Before performing RFR, it is essential to evaluate the assumptions of RFR to ensure the validity of subsequent analysis [14]:

1. **No Missing Value:** Missing value was checked during data preprocessing, and it proved the absence of missing value.

2. **Independence of Observations:** Each row in the Mental Health and Lifestyle Habits dataset represents one country, it proves the uniqueness of each record even the country feature was dropped afterward. The target variable (Happiness Score) is measured per country. Also, in Figure 10.4, it shows the result of the index plot of happiness score and can observe that the happiness scores are randomly scattered across the observation index, this proves that throughout the dataset there is no systematic trend, cyclic pattern, or clustering being observed. Independent observation was represented by each of the data points appearing.

3. **Continuous Numeric Target Variable:** The target variable (Happiness Score) is continuous and numerical. In addition, Figure 10.6 shows the output of the density plot of happiness score, where we can observed that the density plot showing a smooth and continuous distribution of the happiness score across its range, therefore it proved that the happiness score is continuous and numerical. Further proving the continuity, the happiness score was distributed nearly between 1 and 10, without any discrete jumps and gaps.

4. **Sufficient Data Size:** Random Forest relies on building many decision trees, so this dataset contains 3000 rows of records provides enough variability to learn stable and generalizable patterns without overfitting.

5. **Absence of outliers:** After applying the IQR method, all extreme outliers are capped. So, this assumption is met.

```r
# RFR Model Assumption 2

ggplot(mental_health, aes(x = seq_along(Happiness.Score), y = Happiness.Score)) +
  geom_point(alpha = 0.6, color = "darkblue") +
  theme_minimal() +
  labs(
    title = "Index Plot of Happiness Score",
    x = "Observation Index",
    y = "Happiness Score"
  )
```

Figure 10.3 Code of checking the Assumption 2 by plotting Index Plot of Happiness Score

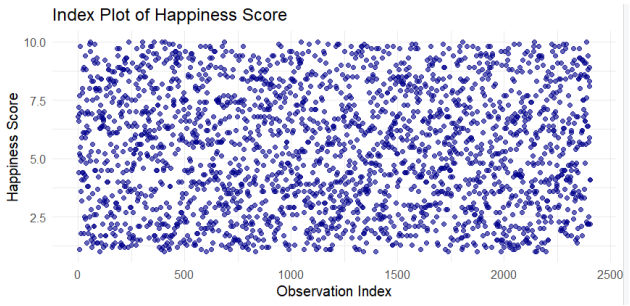

Figure 10.4 Result of the Index Plot of Happiness Score

```r
# RFR Model Assumption 3

ggplot(mental_health, aes(x = Happiness.Score)) +
  geom_density(fill = "skyblue", alpha = 0.6) +
  theme_minimal() +
  labs(
    title = "Density Plot of Happiness Score",
    x = "Happiness Score",
    y = "Density"
  )
```

Figure 10.5 Code of checking the Assumption 3 by plotting Density Plot of Happiness Score

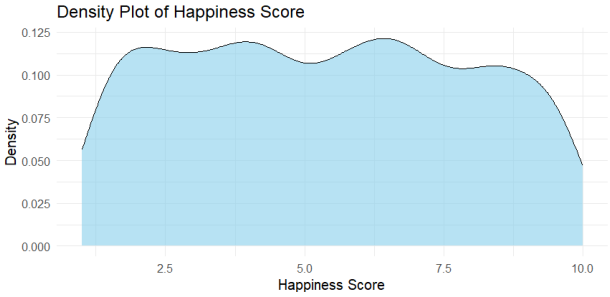

Figure 10.6 Result of the Density Plot of Happiness Score

Considering that all key assumptions are met and knowing the potential non-linear effects of lifestyle habits on mental health, Random Forest Regression is an effective method for modeling happiness scores. Its ability to uncover complex patterns provides meaningful insights into how lifestyle factors collectively influence mental well-being.

B. Perform Regression Analysis

```r
#Random Forest Regression
train_data_rf <- train_data[, !names(train_data) %in% "Happiness.Level"]
test_data_rf  <- test_data[, !names(test_data) %in% "Happiness.Level"]

set.seed(123)
rf_model <- randomForest(
  Happiness.Score ~ .,
  data = train_data_rf,
  ntree = 500,
  mtry = 3,
  importance = TRUE,
)

print(rf_model)
```

Figure 10.7 Code of Random Forest Regression

Before training, the variable Happiness.Level was explicitly removed from the dataset to prevent data leakage. The model was trained with 500 decision trees (ntree=500) to maximize stability and with 3 variables sampled at each split (mtry=3) to ensure diverse feature selection. A random seed was set to guarantee reproducibility of the results. The model also calculated variable importance to show the contribution of each lifestyle habit to the prediction.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

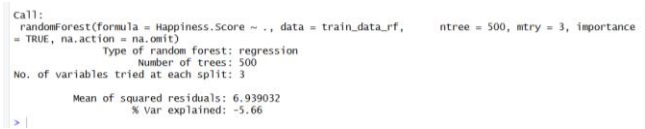

Figure 10.8 Result of Random Forest Regression

Figure 10.8 shows the model has resulted in a Mean of Squared Residuals (MSE) of 6.94 and Variance Explained of -5.66%. The observed negative value indicates that the Random Forest model performs worse than random guessing. The model attempted to learn patterns from the training data, but because those patterns were likely random noise so the model failed to generalize to the test data [15].

Additionally, the MSE of 6.94 translates to a Root Mean Squared Error (RMSE) of approximately 2.63. On a scale of 1 to 10, an average error of 2.63 points is significant. For example, the model might predict a "Happy" person (Score 8) as "Unhappy" (Score 5.4) that makes it statistically unreliable for prediction.

XI. MODEL EVALUATION

```
#Random Forest Evaluation
# Predict on test set
rf_predictions <- predict(rf_model, newdata = test_data_rf)

# Calculate evaluation metrics
MAE_rf  <- mae(test_data_rf$Happiness.Score, rf_predictions)
MSE_rf  <- mse(test_data_rf$Happiness.Score, rf_predictions)
RMSE_rf <- rmse(test_data_rf$Happiness.Score, rf_predictions)

# R-squared
R2_rf <- R2(rf_predictions, test_data_rf$Happiness.Score)

rf_results <- data.frame(
  Metric = c("MAE", "MSE", "RMSE", "R²"),
  Value  = c(MAE_rf, MSE_rf, RMSE_rf, R2_rf)
)

print(rf_results)
```

Figure 11.1 Model Evaluation

The Random Forest model was evaluated on the test dataset using regression metrics such as Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and R-squared (R^2). These metrics provide insight into the model's predictive performance and reliability in estimating happiness levels.

MAE, MSE, and RMSE measure the average prediction error, with lower values indicating more accurate predictions. R-squared indicated how well the model explains the variability in happiness scores, with higher values representing a better fit [16].

The model's predictions on the test dataset are stored in rf_predictions, while the actual Happiness Scores from the test set are in test_data_rf$Happiness.Score.

```
> print(rf_results)
  Metric        Value
1    MAE 2.238887e+00
2    MSE 6.748287e+00
3   RMSE 2.597746e+00
4     R² 3.552299e-05
```

Figure 11.2 Result of Model Evaluation

From Figure 11.2, the MAE is 2.2389, indicating that, on average, the predicted happiness score differs from the actual scores by about 2.24 points. The MSE is 6.7483, and the RMSE is 2.5977, reflecting the magnitude of prediction errors in original and squared units, respectively. The R^2 value is extremely low at 0.00004, suggesting that the model explains almost none of the variability in happiness scores. For information, there is no any detail regarding the adjusted R^2 due to it being only applicable to parametric linear regression models.

The results show that the model's predictions are not closely aligned with the actual value. Although the Random Forest model produces numeric predictions, it failed to capture meaningful patterns in the data. This suggests that the features used are insufficient for predicting happiness, or the model requires further tuning or additional relevant variables to improve accuracy.

However, even though the result failed to capture meaningful patterns in the data, the benefits of using RFR was the ability to model the relationship that consists of a non-linear relationship between the independent variables with the dependent variable which is the happiness score.

XII. CONCLUSION AND FUTURE WORK

In conclusion, throughout the entire research paper, we have performed exploratory data analysis (EDA), hypothesis testing, and regression analysis to identify and proof whether there is a statistically significant relationship between the adults' lifestyle habits factors including the exercise level, diet type, sleep hours, work hours per week, screen time per day, and social interaction score, with the dependent variable which is mental health outcome, assumed as happiness score feature.

In the EDA that has been carried out, we found that all the numerical features do not have any strong relationship with others features, which is observed in the scatter plot and correlation heatmap. Additionally, histogram was plotted as well and we found that excepting from age, others single numerical variables do not have a consistent and stable distribution of the data which means the values are probably variable. Furthermore, we conducted outlier checking in the VII. Data Preprocessing subsection where we plotted a box plot diagram and handled it to remain the reliability of the data.

After conducting the statistical testing, most of the hypotheses we formulated were failed to reject the null hypothesis, which means that overall there is no meaningful relationship between each independent variable with the

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

happiness score. Moreover, in the regression analysis and model evaluation, our selected model which is Random Forest Regression failed to learn the patterns from the training data and the predicted value was not related with the actual data.

Overall, based on the analysis we have completed, we can conclude from our research paper that there is no statistically significant relationship between adults' lifestyle habits with their mental health outcomes. Generally, there is no relationship with these two factors when the dataset may be noisy due to self-reported survey and sampling bias which may reduce the accuracy. However, it is possible there is a relationship between these two factors when changing another more comprehensive and high-quality dataset. This is because the conclusion was drawn based on the dataset we have analyzed. Since this is a mental health topic, which is very psychological and can vary from person to person, and even the lifestyle habits were very complex and may affect the happiness score indirectly. Last but not least, the performance and accuracy of the model can be improved by having more reliable and appropriate features, or even better feature engineering or hyperparameter tuning.

REFERENCES

[1] A. Soundankar, “Mental Health and Lifestyle Habits (2019–2024),” *Kaggle*, Mar. 17, 2025. [Online]. Available: https://www.kaggle.com/datasets/atharvasoundankar/mental-health-and-lifestyle-habits-2019-2024

[2] E. P. Terlizzi and B. Zablotsky, *Mental Health Treatment Among Adults: United States, 2019*. Hyattsville, MD: National Center for Health Statistics, 2020.

[3] H. Herrman and E. Jané-Llopis, “The Status of Mental Health Promotion,” *Public Health Reviews*, vol. 34, no. 2, pp. 1–2, Dec. 2012.

[4] World Health Organization, “Suicide,” *World Health Organization*, Mar. 25, 2025. https://www.who.int/news-room/fact-sheets/detail/suicide

[5] M. J. Lay San Too, “The association between mental disorders and suicide: A systematic review and meta-analysis of record linkage studies,” *Journal of Affective Disorders*, vol. 312, 2019.

[6] H. Y. An, W. Chen, C. W. Wang, H. F. Yang, W. T. Huang, and S. Y. Fan, “The relationships between physical activity and life satisfaction and happiness among young, middle-aged, and older adults,” *International Journal of Environmental Research and Public Health*, vol. 17, no. 13, p. 4817, 2020. Available: https://doi.org/10.3390/ijerph17134817

[7] H. Herrman, S. Saxena, and R. Moodie, “Promoting mental health: Concepts, emerging evidence, practice,” *Public Health Reviews*, vol. 34, no. 2, 2012. Available: https://doi.org/10.1007/BF03391698

[8] A. H. S. Ghahfarokhi *et al.*, “The association between dietary behavior, diet quality, lifestyle scores and happiness levels,” *BMC Nutrition*, vol. 10, p. 45, 2024. Available: https://doi.org/10.1186/s40795-024-00735-4

[9] IBM, “Exploratory Data Analysis,” *Ibm.com*. https://www.ibm.com/think/topics/exploratory-data-analysis

[10] GeeksforGeeks, “What is Exploratory Data Analysis?,” *GeeksforGeeks*, Jul. 22, 2021. https://www.geeksforgeeks.org/data-analysis/what-is-exploratory-data-analysis/

[11] S. Nickolas, “What does it mean if the correlation coefficient is positive, negative, or zero?,” *Investopedia*, Dec. 26, 2024. https://www.investopedia.com/ask/answers/032515/what-does-it-mean-if-correlation-coefficient-positive-negative-or-zero.asp

[12] R. W. Nahhas, Introduction to Regression Methods for Public Health using R. 1st ed. Chapman and Hall, 2024. Available: https://www.bookdown.org/rwnahhas/RMPH/mlr-collinearity.html

[13] M. J. Diamantopoulou, R. Özçelik, and Ş. K. Genç, “Evaluation of the random forest regression machine learning technique as an alternative to ecoregional based regression taper modelling,” *Computers and Electronics in Agriculture*, vol. 239, pt. A, 2025, Art. no. 110964. doi: 10.1016/j.compag.2025.110964.

[14] J. Ehrlinger, “ggRandomForests: Visually exploring a Random Forest for regression,” *arXiv preprint arXiv:1501.07196*, 2015. [Online]. Available: https://arxiv.org/abs/1501.07196

[15] A. Liaw and M. Wiener, “Classification and Regression by randomForest,” *R News*, vol. 2, no. 3, pp. 18–22, Dec. 2002. Available: https://journal.r-project.org/articles/RN-2002-022/RN-2002-022.pdf

[16] A. Botchkarev, "A new typology design of performance metrics to measure errors in machine learning regression algorithms," *Interdisciplinary Journal of Information, Knowledge, and Management*, vol. 14, pp. 45–76, 2019.