A PROJECT REPORT
on
EXPLORING THE IMPACT OF SOCIAL MEDIA ENGAGEMENT ON THE MENTAL WELLBEING OF UNIVERSITY STUDENTS USING DASS-21 QUESTIONNAIRE AND R-PROGRAMMING

SUBMITTED IN PARTIAL FULFILMENT OF THE REQUIREMENT FOR THE AWARD OF THE DEGREE OF

Bachelor Of Science
In
Statistics

SUBMITTED BY:

SHUBHAM TEOTIA	21STBSA183
ANZAR KHAN	21STBSA238
MD AABISH RAHMAN	21STBSA350
MOHAMMAD SALMAN FURQAN	21STBSA361
ARBAZ KHAN	21STBSA379
ARHAM ABBASI	21STBSA383
ARSALAN AHMAD FAROOQUI	21STBSA395
HUSSAIN AMAAN ASHIQUE	21STBSA411
MOHAMMAD WALI	21STBSA439
SAYED SHUMAYEL ALI	21STBSA457
DILSHAD AHMAD	21STBSA471
MOHAMMAD NAFEES IQBAL	21STBSA924

Under the Supervision of
PROF. AQUIL AHMED
&
DR. AIJAZ AHMAD DAR

DEPARTMENT OF STATISTICS & OPERATIONS RESEARCH
ALIGARH MUSLIM UNIVERSITY ALIGARH
ALIGARH (202002), U.P., INDIA

2024

DEPARTMENT OF STATISTICS AND OPERATIONS RESEARCH
ALIGARH MUSLIM UNIVERSITY
ALIGARH (202002), U.P., INDIA

CERTIFICATE

This is to certify that the project titled “EXPLORING THE IMPACT OF SOCIAL MEDIA ENGAGEMENT ON THE MENTAL WELLBEING OF UNIVERSITY STUDENTS USING DASS21 QUESTIONNAIRE AND R-PROGRAMMING” has been carried out by the following group of students of Bachelor’s of Science (Statistics), Final Year, 6th Semester, 2023-24, under my supervision for the course “STB6S1Project”.

SUBMITTED BY:

SHUBHAM TEOTIA	21STBSA183
ANZAR KHAN	21STBSA238
MD AABISH RAHMAN	21STBSA350
MOHAMMAD SALMAN FURQAN	21STBSA361
ARBAZ KHAN	21STBSA379
ARHAM ABBASI	21STBSA383
ARSALAN AHMAD FAROOQUI	21STBSA395
HUSSAIN AMAAN ASHIQUE	21STBSA411
MOHAMMAD WALI	21STBSA439
SAYED SHUMAYEL ALI	21STBSA457
DILSHAD AHMAD	21STBSA471
MOHAMMAD NAFEES IQBAL	21STBSA924

Prof. Aquil Ahmed
(Supervisor)

(Chairman)

Dr. Aijaz Ahmad Dar
(Supervisor)

ACKNOWLEDGEMENT

In the Name of **Almighty**, the Most Gracious, the Most Merciful.

All praise and thanks are due to Almighty, the Lord of the Worlds, for His blessings and guidance throughout the completion of this project. We thank Him for His infinite mercy and grace.

We would also like to express our deep gratitude to our supervisors, Prof. Aquil Ahmed (Chairman) & Dr. Aijaz Ahmad Dar (Assistant Professor), Department of Statistics & Operations Research, Aligarh Muslim University, Aligarh for their constant guidance, valuable insights, and support throughout the project. Their expert advice and encouragement have been invaluable to our success.

We also extend our heartfelt thanks to the Chairman of the Department for providing us with all the necessary resources and facilities to carry out our project work.

We are also thankful to the each one of those who undertake the survey without which it was not possible to carry out this piece of work. Furthermore, we would like to thank our parents, siblings, and colleagues for their unwavering support and encouragement throughout this journey. Their love and encouragement were the driving forces behind our success.

We would express our sincere appreciation to all those who have contributed to this project and helped us in one way or another. May the **Almighty** reward them abundantly.

Dedicated to

Late Sir Syed Ahmed Khan

(The Founder of Aligarh Muslim University, Aligarh),

Our Parents & Professors

CONTENTS

Chapter I
Introduction

[1]1.1 Introduction

The rapid proliferation of social media platforms has transformed the way we interact, communicate, and share information. With over 4.2 billion active users worldwide, social media has become an indispensable part of modern life. However, the increasing use of social media has also raised concerns about its impact on mental health, particularly among young people. The constant stream of information, the pressure to present a perfect online image, and the fear of missing out (FOMO) can all contribute to feelings of anxiety, depression, and loneliness.

The relationship between social media use and mental health is complex and multifaceted. While social media can provide a sense of connection and community, it can also exacerbate feelings of isolation and disconnection. The constant comparison to others, the fear of being left out, and the pressure to present a perfect online image can all take a toll on mental health. Furthermore, the lack of face-to-face interaction and the potential for cyberbullying can also have negative effects on mental well-being.

Despite the growing concern about the impact of social media on mental health, there is a need for more research in this area. While some studies have suggested a positive correlation between social media use and mental health outcomes, others have found no significant relationship. The inconsistent findings highlight the need for further investigation into the complex interplay between social media use and mental health.

This project aims to contribute to the existing body of research by investigating the relationship between social media usage and various mental health outcomes, including depression, anxiety, and overall well-being. The study will focus on a sample of University students, who are among the most active users of social media. By analysing data collected from a survey questionnaire, this project will explore the relationships between social media usage patterns and mental health outcomes.

Social Media

---

[1] The Content of this chapter is prepared with the help of various resources whose references are given the references section.

Social media refers to a range of digital platforms and applications that enable users to create, share, and engage with content, as well as to connect and interact with others online. These platforms have become ubiquitous in modern society, with a significant portion of the global population actively using various social media platforms on a regular basis.
In the context of this project, which aims to investigate the impact of social media on the mental health of university students, it is essential to provide a comprehensive understanding of the different aspects of social media and its role in the lives of young people.

The Role of Social Media in the Lives of University Students

University students, as a demographic, have a unique relationship with social media. This population is often considered digital natives, having grown up in an era where social media platforms have become an integral part of their daily lives. Some key aspects of the role of social media in the lives of university students include:

Social Connections and Relationships:

Social media platforms play a significant role in the way university students maintain and cultivate their social connections, both with their peers and with their broader social networks. These platforms provide a virtual space for students to stay connected, share experiences, and engage in ongoing communication, even when physical distance separates them. Social media can help students expand their social circles, reconnect with old friends, and foster new relationships, which can be particularly valuable during the transitional period of university life.

Identity Formation and Self-Presentation:

The ability to curate and present an online persona can be particularly influential during the formative years of university education, as students navigate the process of self-discovery and identity development. Social media allows students to experiment with different aspects of their identity, explore their interests and values, and present a carefully crafted image to their online audience. This process of self-expression and self-representation can be a crucial part of the university experience, as students strive to establish their unique identities and find their place within the larger social landscape.

Information Sharing and Learning:

Social media can serve as a platform for the exchange of information, ideas, and knowledge, which can be particularly valuable for university students in their academic and personal pursuits. Students can use social media to share and access educational resources, engage in discussions with their peers, and stay informed about campus events, academic opportunities, and extracurricular activities. This can enhance their learning experience, foster intellectual discourse, and facilitate the acquisition of knowledge beyond the classroom.

Entertainment :

Social media can also provide a source of entertainment and leisure for university students, offering a means of relaxation and distraction from the demands of academic life. Students may use social media platforms to consume and share content, such as videos, memes, and viral trends, as a way to unwind and recharge. This can help students maintain a healthy balance between their academic responsibilities and their personal well-being, providing a much-needed outlet for stress relief and social engagement.

Mental Health and Well-Being:

The impact of social media on the mental health and well-being of university students is a central focus of this project. The complex interplay between social media usage and mental health outcomes is a topic of growing concern, as research has suggested both positive and negative effects. While social media can facilitate social connections and provide a sense of belonging, it can also contribute to feelings of anxiety, depression, and loneliness, particularly when used in an unhealthy or excessive manner. Understanding the nuances of this relationship is crucial for supporting the overall well-being of university students.

Social Media:

Social media is the collective of online communications channels dedicated to community-based input, interaction, content-sharing and collaboration Websites and applications dedicated to forums, microblogging, social networking, social bookmarking, social curation, and wikis are among the different types of social media.

Commonly Used Social Media Platforms

Facebook: One of the most widely used social networking platforms, plays a crucial role in the lives of university students. It serves as a virtual space for students to connect with classmates, join university groups, share academic resources, and stay updated on campus events. Facebook's features, such as event invitations, group discussions, and messaging, facilitate communication and collaboration among students, making it a central hub for social interaction and information sharing within the university community.

Snapchat: Snapchat's unique features, including disappearing messages, stories, and filters, make it a popular platform among university students for casual and creative communication. Students use Snapchat to share real-time moments, express themselves through multimedia content, and maintain a sense of immediacy and authenticity in their interactions. The platform's emphasis on visual communication and temporary content fosters a dynamic and engaging social experience for students.

Telegram: Known for its secure messaging and file-sharing capabilities, provides a reliable platform for university students to communicate privately and efficiently. Students use Telegram for group chats, project collaborations, sharing academic resources, and organizing study sessions. The platform's encryption features ensure data privacy and security, making it a preferred choice for academic and professional communication among students.

Twitter (X): With its real-time updates, character limit, and hashtag system, serves as a valuable social networking tool for university students to engage in public conversations, share insights, and stay informed about current events and trends. Students use Twitter to follow academic influencers, participate in discussions on various topics, and express their opinions concisely. The platform's fast-paced nature and diverse content make it a dynamic source of information and interaction for students.

Instagram: A visual-centric platform, allows university students to share photos, videos, and stories that reflect their experiences, interests, and creativity. Students use Instagram to showcase their personal lives, hobbies, and achievements, creating a visual narrative of their university journey. The platform's emphasis on aesthetics, storytelling, and community engagement enables students to express themselves artistically, connect with like-minded individuals, and build a digital portfolio of their university experiences.

WhatsApp: A widely used messaging app, plays a vital role in the communication and coordination efforts of university students. Students use WhatsApp for one-on-one chats, group discussions, file sharing, and voice/video calls, making it a versatile platform for staying connected with peers, organizing study groups, and coordinating academic activities.

Social media platforms can be categorized into several broad types, each with its unique features and functionalities:

1. **Social Networking Platforms**:
   - o Platforms such as Facebook, Twitter, and LinkedIn allow users to create personal profiles, connect with friends and followers, and share a wide range of content, including text, images, videos, and links.
2. **Microblogging Platforms**:
   - o Platforms like Twitter and Tumblr enable users to post short, concise updates or "microblog" entries, often accompanied by hashtags and links.
3. **Image and Video Sharing Platforms**:
   - o Platforms such as Instagram, Snapchat, and TikTok focus on the sharing and consumption of visual content, including photographs, short videos, and animated clips.
4. **Messaging and Instant Messaging Platforms**:
   - o Apps like WhatsApp, Facebook Messenger, and iMessage allow users to engage in real-time, one-on-one or group-based communication through text, voice, and video messages.
5. **Discussion Forums and Online Communities**:
   - o Platforms like Reddit, Discord, and various online forums provide spaces for users to engage in topical discussions, share knowledge, and participate in niche communities.
6. **Collaborative and Content-Sharing Platforms**:
   - o Platforms like YouTube, Twitch, and Wikipedia enable users to create, share, and consume a wide range of user-generated content, from videos and live streams to collaborative encyclopedic entries.

Chapter II
Research Methodology

[2]2.0 Objective

To systematically investigate and analyze the impact of social media usage on the mental health of university students, with the aim of identifying patterns, correlations, and potential causative factors that contribute to their psychological well-being.

2.1.1 Data Collection

The DASS-21 (Depression, Anxiety, and Stress Scales - 21 items) is a widely used self-report questionnaire designed to measure the severity of symptoms related to depression, anxiety, and stress. It is a shorter version of the original DASS, which consists of 42 items. The DASS-21 is considered a reliable and valid tool for assessing emotional states and psychological distress.

Structure of DASS-21 :

The DASS-21 consists of three subscales, each containing 7 items:

Depression: - Items assess dysphoria, hopelessness, devaluation of life, self-deprecation, lack of interest/involvement, anhedonia, and inertia.

Anxiety:- Items assess autonomic arousal, skeletal muscle effects, situational anxiety, and subjective experience of anxious affect.

Stress:- Items assess difficulty relaxing, nervous arousal, and being easily upset/agitated, irritable/over-reactive, and impatient.

Scoring and Interpretation

Participants rate each item on a 4-point scale (0-3) based on how much each statement applied to them over the past week. Scores are then totaled for each subscale, with higher scores indicating higher levels of depression, anxiety, or stress.

Use of DASS-21 Dataset

Researchers use the DASS-21 in various studies to assess psychological distress and emotional states among different populations. Datasets containing responses to the DASS-21 can be valuable for analyzing relationships between mental health and other variables such as demographics, health outcomes, or treatment interventions.

---

[2] The Content of this chapter is prepared with the help of various resources whose references are given the references section.

Accessing DASS-21 Dataset

To access a specific dataset containing DASS-21 responses, you can search academic databases, research repositories, or contact researchers who have used this questionnaire in their studies. Datasets associated with the DASS-21 may be available for secondary analysis or replication studies, subject to data sharing agreements and ethical considerations.

When using or analyzing DASS-21 datasets, it's essential to consider the context of the study, population demographics, and any limitations or biases inherent in self-report measures. Proper acknowledgment and citation of the original authors and dataset sources are also important for transparency and academic integrity.2.3. Graphical Presentation.

2.1.2 Target Population & Sample Size

The dataset focuses on university students as the target population, providing a comprehensive view of various factors that may influence their well-being and academic experiences. The inclusion of demographic variables such as Gender and Level of Study allows for comparative analyses across different student subgroups. The Area of Study variable enables researchers to investigate potential differences in screen time, social media usage, and mental health indicators among students from diverse academic disciplines.

Key Variables:

I. Gender: This binary variable (male/female) enables researchers to examine gender-based differences in screen time, social media usage, and mental health outcomes among university students.

II. Level of Study: This ordinal variable (UG, PG, Phd) allows for the exploration of potential differences in the studied factors across undergraduate, postgraduate, and doctoral students, providing insights into how academic level may influence the variables of interest.

III. Area of Study: This nominal variable (Science, Engineering, Medicine, Arts, Commerce) facilitates the investigation of potential differences in screen time, social media usage, and mental health indicators among students from various academic fields, enabling researchers to identify discipline-specific trends and patterns.

IV. Total Screen Time: This continuous variable, measured in minutes, serves as a key predictor variable, allowing researchers to assess the impact of screen time on students' mental health and well-being.

V. Total Social Media Usage Time: Similar to Total Screen Time, this continuous variable, also measured in minutes, enables researchers to examine the influence of social media usage on university students' mental health and academic experiences.
VI. Stress Score, Depression Score, Anxiety Score: These continuous variables, measured on a scale, provide valuable insights into students' mental health status, serving as both predictor and outcome variables in the analysis. Researchers can investigate the relationships between these mental health indicators, screen time, social media usage, and other factors.
VII. Experience of Cyberbullying: This nominal variable (yes/no/maybe) allows for the examination of the prevalence and impact of cyberbullying among university students, shedding light on the challenges they face in the digital realm and its potential influence on their mental health and well-being.

Significance and Implications:

By analyzing this comprehensive dataset, researchers can gain valuable insights into the complex interplay between demographic factors, academic experiences, screen time, social media usage, and mental health indicators among university students. The findings can inform targeted interventions, support services, and policy decisions aimed at promoting student well-being and academic success in the digital age.

2.2 Data analysis tools

2.2.1 R-programming

R is an open-source programming language designed primarily for statistical computing and data analysis. Here are some key points about R:

- Definition: According to the R Foundation, R is "a language and environment for statistical computing and graphics."
- Data Analysis Tool: R is widely used as a data analysis and statistical software tool.
- Features:
    - High-performance data storage and handling facilities.
    - Operators for array calculations (especially matrices).
    - Integrated tools for data analysis and graphical display.
    - Simple and effective programming language with user-defined functions, loops, conditionals, and I/O facilities.

Advantages of R:

- Open-Source: R is freely available, making it a low-risk choice for developers.
- Platform-Independent: R runs on Windows, Macintosh, UNIX, and Linux platforms, allowing developers to create one program that works across different systems.
- Rich Package Ecosystem: R has an extensive collection of packages (more than 10,000 in the CRAN repository) that provide reusable code, documentation, and sample data.
- Applications in Diverse Fields:
    - Academics: R is widely used in research and academic settings.
    - Healthcare: For statistical analysis of medical data.
    - Government: For policy analysis and decision-making.
    - Finance: Risk modeling, portfolio analysis, and econometrics.
    - Retail: Customer segmentation, sales forecasting, and inventory management.
    - Media: Sentiment analysis, recommendation systems, and content optimization.
    - Manufacturing: Quality control, process optimization, and supply chain analytics.
    - Technology: Web analytics, A/B testing, and performance monitoring.
    - Energy: Renewable energy modeling and optimization.
    - Bioinformatics: Genomic data analysis and drug discovery.
    - And More!

In summary, R's versatility, rich ecosystem, and widespread adoption across various domains make it a powerful tool for data scientists, statisticians, and researchers.

2.3 Data Preparation

Upon receiving the survey responses from the college students, the data underwent a series of steps to ensure its accuracy and readiness for analysis. The first task was to import the data into R.

The data was carefully checked for any missing values or inconsistencies in the data. Some participants had left certain questions unanswered, resulting in missing data points. To address this, incomplete responses were excluded from the analysis to maintain the integrity of the data.

The continuous variables, like total screen time, total social media usage time, and the DASS-21 scores for depression, anxiety, and stress, were assessed for normality using the Shapiro-Wilk test. It was found that these variables were not normally distributed. To prepare them for analysis, a Box-Cox transformation was applied to normalize the data.

The categorical variables, including gender, level of study, area of study, and cyberbullying experience, were coded as 0, 1, 2, 3, or 0, 1, depending on the variable.
After transforming and cleaning the data, a structured and organized dataset was created, ready for statistical analysis. This dataset was saved in a CSV format suitable for R.

2.4 Statistical Analysis

Statistical analysis is a fundamental component of data analysis that involves the application of statistical methods to interpret, summarize, and draw meaningful conclusions from data. It encompasses a range of techniques and procedures used to explore relationships, patterns, and trends within datasets, providing valuable insights into the underlying structure and characteristics of the data.

2.4.1 Descriptive Statistics

Descriptive statistics involve summarizing and organizing data characteristics, providing simple summaries about the sample and the measures. These statistics are crucial in quantitative analysis, forming the basis for understanding data through measures like central tendency (mean, median, mode) and dispersion (range, variance, standard deviation). They help simplify large datasets into manageable summaries, aiding in data interpretation and pattern identification. Descriptive statistics are essential for summarizing data effectively, enabling comparisons and facilitating informed decision-making processes based on the data's key features and trends.

2.4.2 Normality Testing

Normality testing is a crucial step in data analysis that evaluates whether a dataset follows a normal distribution. A normal distribution is a fundamental assumption for many statistical tests, and regression analysis . If the data are not normally distributed, the validity and interpretation of these tests may be compromised.

The Shapiro-Wilk test is a widely recommended normality test that provides better power than other methods like the Kolmogorov-Smirnov test. It assesses the null hypothesis that the data are normally distributed by comparing the sample data to a normal distribution. The test statistic, W, ranges from 0 to 1, with values closer to 1 indicating a normal distribution.

2.4.3 Graphical Presentation

The research methodology employed in this study investigating the impact of social media on mental health and cyberbullying among college students is graphically depicted through a series of visualizations. These graphical representations serve to elucidate the research process, sample characteristics, and analytical procedures in a clear and concise manner.

2.4.3.1 Bar Chart

A bar chart, also known as a bar graph, is a graphical representation of data using rectangular bars. The length or height of each bar is proportional to the value it represents, and the bars can be oriented either horizontally or vertically. The x-axis represents the categories being compared, while the y-axis represents the scale or value of the data being measured.

Bar charts are commonly used to compare the values of different categories or groups, as well as to display trends over time. They are often used in business, finance, economics, and other fields to visualize data and make it easier to understand.

Bar charts can be created using various software programs, including Microsoft Excel, Google Sheets, and other data visualization tools. They can be customized with different colors, fonts, labels, and other design elements to make the data more visually appealing and easier to interpret.

2.4.3.2 Pie Chart

A pie chart is a circular statistical graphic that is divided into slices to illustrate numerical proportion. Each slice represents a category, and the area of the slice is proportional to the quantity it represents. Pie charts are commonly used to show the proportional sizes of different parts of a whole. They are effective for visualizing the relative sizes of categories within a dataset. However, pie charts are limited in their ability to display precise numerical values and should be used judiciously, as they can be misleading if the categories do not sum to a meaningful whole.

2.3.4.4 Histogram

A histogram is a graphical representation that organizes a group of data points into user-specified ranges. It is an estimate of the probability distribution of a continuous variable. The histogram is constructed by first dividing the range of values in the data into a series of contiguous intervals, called bins or classes. Then, for each interval, a rectangle is drawn with a base width equal to the interval range and an area proportional to the frequency of values in that interval.

2.3.4.5 Scatter Plot

A scatter plot, also known as a scattergram or scatter chart, is a graphical representation of the relationship between two variables. It displays the values of two variables as points on a two-dimensional plane, with the horizontal axis representing one variable and the vertical axis representing the other.

2.3.4.6 Heat Map

Heat maps, also known as heatmaps, are 2-dimensional data visualization techniques that represent the magnitude of individual values within a dataset using colors. They are used to display data patterns, relationships, and distributions visually.

2.4.4 Correlation Analysis

Correlation analysis is a statistical technique used to measure the strength and direction of the linear relationship between two variables. It quantifies the degree to which changes in one variable are associated with changes in another variable. The correlation coefficient, denoted as r, ranges from -1 to 1, with -1 indicating a perfect negative linear relationship, 0 indicating no linear relationship, and 1 indicating a perfect positive linear relationship.

A correlation matrix is a table that displays the correlation coefficients for different variables in a dataset. It is used to explore the relationship between variables and identify patterns or trends. Correlation coefficients measure the strength and direction of the linear relationship between two variables. The correlation coefficient ranges from -1 to +1. A coefficient of +1 indicates a perfect positive correlation; a coefficient of 0 indicates no correlation, and a coefficient of -1 indicates a perfect negative correlation.

2.4 Statistical Tests

2.4.1 Proportion Test

A proportion test is a statistical method used to determine whether the proportion of a particular characteristic or attribute in a sample is significantly different from a known or hypothesized proportion in the population. The proportion test is a hypothesis testing technique that involves comparing the observed proportion in the sample with the hypothesized proportion in the population, using a statistical test such as the z-test or chi-square test.

A proportion test involves comparing two sample proportions to determine if they are statistically different from each other. The null hypothesis in a proportion test is that the two proportions are equal, while the alternative hypothesis is that they are not equal. A test statistic, such as the z-score or chi-square statistic, is used to calculate the probability of obtaining the observed difference in proportions if the null hypothesis is true. If the probability, or p-value, is below a certain threshold, typically 0.05 or 0.01, then the null hypothesis is rejected, and it is concluded that there is a significant difference between the two proportions.

There are several types of proportion tests that can be used depending on the nature of the data and research question:

- **One-sample proportion test:** compares a sample proportion to a known population proportion or a hypothesized proportion
- **Two-sample proportion test:** compares two sample proportions from independent samples
- **Paired proportion test:** compares two sample proportions from paired or matched data, such as before-and-after measurements on the same individuals.
- **Goodness-of-fit proportion test:** tests whether observed proportions in a sample follow an expected distribution or proportion, such as a binomial or normal distribution.

Assumptions of proportion tests

Proportion tests rely on several assumptions, including:

i. The data are binary or dichotomous, meaning there are only two possible outcomes.
ii. The samples are random and independent.
iii. The sample sizes are large enough to assume a normal distribution of the sample proportion, typically at least 10 successes and 10 failures.
iv. The null hypothesis is specified in advance, and the test is appropriately powered to detect a meaningful difference between the proportions, if one exists.

2.4.2 Chi-Square Test

The chi-square test is a statistical method used to determine whether there is a significant association between two categorical variables. It is a non-parametric test, which means that it does not require any assumptions about the distribution of the data being analyzed. This test is widely used in social and health sciences, as well as market research and quality control. The chi-square test is based on the comparison of the observed frequencies of different categories with the expected frequencies, assuming that there is no association between the variables. If the observed frequencies differ significantly from the expected frequencies, then the chi-square test indicates that there is a significant association between the variables. A chi-square test is a valuable tool in research and data analysis, providing insights into the relationships between different categorical variables.

The chi-square test is set up by first defining the null hypothesis and the alternative hypothesis. The null hypothesis states that there is no association between the two categorical variables being compared, while the alternative hypothesis states that there is a significant association between the variables.

Next, a contingency table is created, which shows the observed frequencies of the different categories for each of the variables being analyzed. The contingency table is a two-way table, with the rows representing one variable and the columns representing the other variable. Once the contingency table has been created, the expected frequencies are calculated for each cell of the table, assuming that there is no association between the variables. The expected frequency for each cell is calculated by multiplying the marginal totals for the row and column that the cell belongs to and then dividing by the total sample size.

The chi-square statistic is then calculated by summing the squared differences between the observed and expected frequencies for each cell and dividing by the expected frequency for that cell. The resulting value is compared to a chi-square distribution table with degrees of freedom equal to the product of the number of rows minus one and the number of columns minus one. If the calculated chi-square value is greater than the critical value from the chisquare distribution table, then the null hypothesis is rejected, and it is concluded that there is a significant association between the two variables being compared.

Applications of chi-square tests There are several applications of chi-square tests that can be used, depending on the nature of the data and the research question:

- **Chi-square goodness-of-fit test:** compares the observed frequencies of a single categorical variable to the expected frequencies from a specified distribution or proportion.
- **Chi-square test of independence:** compares the observed frequencies of two categorical variables to the expected frequencies, assuming no association between them.
- **Chi-square test for homogeneity:** compares the observed frequencies of a categorical variable across multiple populations or groups, assuming no association between the variable and the groups.

Assumptions of chi-square tests

The chi-square test relies on several assumptions, including:

i. The data are categorical, or count data.
ii. The samples are random and independent.
iii. The expected frequencies are greater than 5 for each cell in the contingency table, or 10 for smaller sample sizes.
iv. The null hypothesis is specified in advance, and the test is appropriately powered to detect a meaningful difference between the observed and expected frequencies.

2.4.3 Linear Regression

Regression is a statistical method used to establish a relationship between a dependent variable and one or more independent variables. In simpler terms, regression analysis helps to find the relationship between a dependent variable (outcome) and one or more independent variables (predictors). Regression analysis can be used to explore and understand the relationships between variables and to predict the future values of the dependent variable based on the values of the independent variable(s). Regression analysis is often used in many fields, including economics, finance, marketing, biology, psychology, and engineering. It is used to investigate the causal relationship between variables and make predictions based on historical data. There are many types of regression models, including linear regression, logistic regression, polynomial regression, and multiple regression. The choice of model depends on the nature of the data and the research question being addressed.

Linear regression is one of the most widely used regression models. It assumes a linear relationship between the dependent variable and the independent variable(s). The goal of linear regression is to find the best-fit line that explains the relationship between the variables. Multiple regression is used when there are two or more independent variables that affect the dependent variable. The goal of multiple regression is to find the best-fit line that explains the relationship between the variables. Regression analysis is a powerful tool for understanding and predicting the relationships between variables.

A linear model specifies a linear relationship between a dependent variable and independent variables.

$$y = \beta 0 + \beta 1 * X1 + \ldots\ldots\ldots\ldots\ldots + \beta n * Xn + \in$$

where y is the dependent variable, {Xi} are independent variables, {βi} are parameters of the model.

Assumptions of Regression Analysis

i. **Linearity**: The relationship between the dependent variable and the independent variable(s) should be linear.
ii. **Independence**: The observations should be independent of each other.
iii. **Homoscedasticity**: The variance of the errors should be constant across the range of the independent variable(s).
iv. **Normality**: The errors should be normally distributed.
v. **No multicollinearity**: The independent variables should not be highly correlated with each other.

2.4.4 Shapiro-Wilk Test

The Shapiro-Wilk test is a statistical test of the hypothesis that the distribution of the data as a whole deviates from a comparable normal distribution. If the test is non-significant (p>. 05), it tells us that the distribution of the sample is not significantly different from a normal distribution.

The Shapiro–Wilk test tests the null hypothesis that a sample x1, ..., xn came from a normally distributed population.
The test statistic is

$$W = \frac{\left(\sum_{i=1}^{n} a_i x_{(i)}\right)^2}{\sum_{i=1}^{n} (x_i - \overline{x})^2},$$

where

- W is the Shapiro-Wilk test statistic,
- n is the sample size,
- $x(i)$ represents the ordered sample values,
- x_i is the i^{th} ordered sample value,
- x^{-} is the sample mean,
- a_i are the constants specific to the sample size and distribution, and are calculated using statistical software.

The null-hypothesis of this test is that the population is normally distributed. Thus, if the p value is less than the chosen alpha level, then the null hypothesis is rejected and there is evidence that the data tested are not normally distributed. On the other hand, if the p value is greater than the chosen alpha level, then the null hypothesis (that the data came from a normally distributed population) can't be rejected (e.g., for an alpha level of .05, a data set with a p value of less than .05 rejects the null hypothesis that the data are from a normally distributed population – consequently, a data set with a p value greater than the .05 alpha value fails to reject the null hypothesis that the data is from a normally distributed population).

2.4.5 Mann-Whitney U test

The Mann-Whitney U test is a non-parametric statistical test used to compare two independent samples or groups. It assesses whether two sampled groups are likely to derive from the same population or have the same shape of data distribution.
The null hypothesis (H0) states that the two populations are equal, while the alternative hypothesis (H1) suggests that the two populations are not equal

The mathematical formula for calculating the Mann-Whitney U test statistic is:

$$U = n_1 n_2 + \frac{n_2(n_2+1)}{2} - R_2$$

Where,
n1 is the sample size of the first group
n2 is the sample size of the second group
R2 is the sum of the ranks in the second group

Chapter III
Statistical Analysis

3.1 Descriptive Statistics

Descriptive statistics play a crucial role in summarizing and understanding the key characteristics of the data collected for the project on the impact of social media on mental health and cyberbullying among college students.

For the categorical variables, including gender, level of study, area of study, and cyberbullying experience, frequency tables and bar plots were generated. These descriptive measures and visualizations help in understanding the distribution of the sample across different categories and identifying any potential patterns or imbalances.

For the continuous variables, such as total screen time, total social media usage time, and the DASS-21 scores for depression, anxiety, and stress, measures of central tendency (median) and dispersion (interquartile range and range) were calculated. These statistics provide insights into the typical values and the spread of the data for each variable.

3.1.1 Gender

The dataset includes 2 categories for Gender: male and female. We can calculate the frequency and percentage of each category to understand the distribution of gender in the sample.

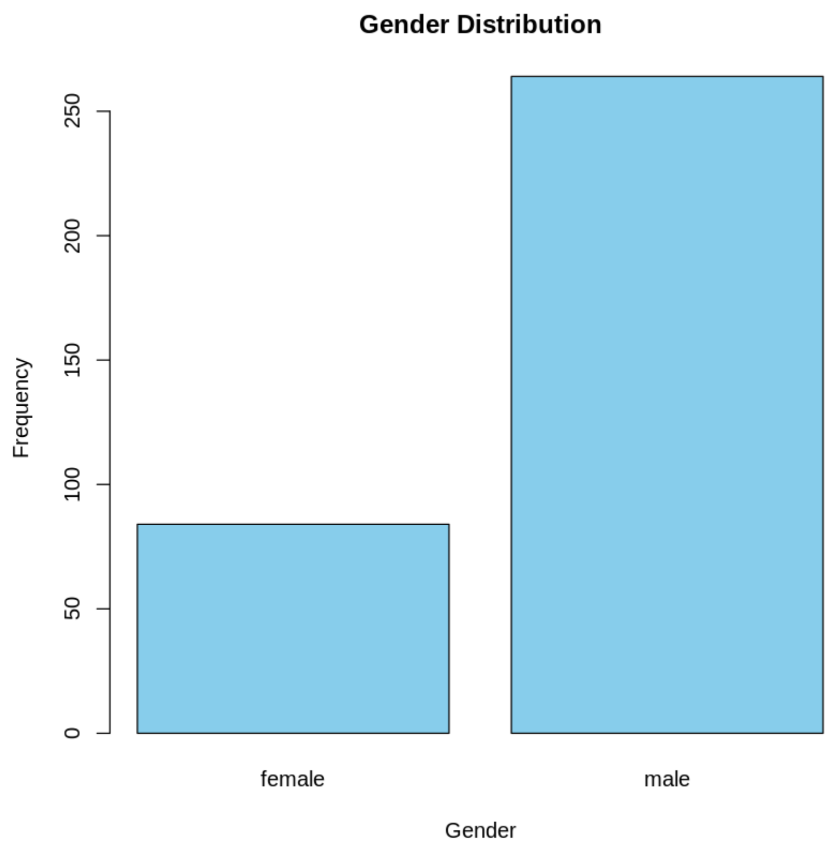
Using R programming to calculate **Frequency**

```
> gender_freq <- table(Gender)
> print (gender_freq)
> gender_percent <- prop.table(gender_freq) * 100
> print(gender_percent)
```

	Female	Male
	84	264
Percentage	23.13	75.86

Using R programming to calculate **Barplot**

> barplot(gender_freq, main = "Gender Distribution", xlab = "Gender", ylab = "Frequency", col = "skyblue")



3.1.2 Level of Study

The dataset includes 3 categories for Level of Study: UG, PG, and PhD. We can calculate the frequency and percentage of each category to understand the distribution of study levels in the sample.

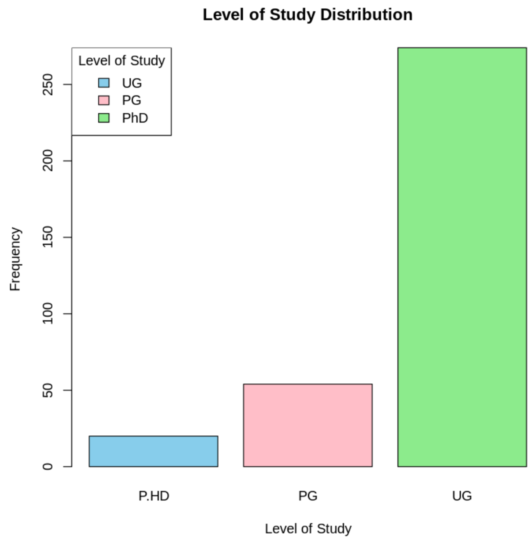
Using R programming to calculate frequency, percentage and bar chart

```
>los_freq <- table(LOS) # Frequency
> study_percent <- prop.table(los_freq) * 100  # Percentage
```

	P.HD	PG	UG
	20	54	274
Percentage	5.74	15.51	78.73

Barplot using R programming

```
> colors <- c("skyblue", "pink", "lightgreen")  # Define colors for each category
> barplot(los_freq, main = "Level of Study Distribution", xlab = "Level of Study", ylab = "Frequency", col = colors)
> legend("topleft", legend = c("UG", "PG", "PhD"), fill = colors, title = "Level of Study")
```



3.1.3 Area of Study

The dataset includes 5 categories for Area of Study: Science, Engineering, Medicine, Arts, and Commerce. We can calculate the frequency and percentage of each category to understand the distribution of study areas in the sample.

Using R programming to calculate frequency& percentage

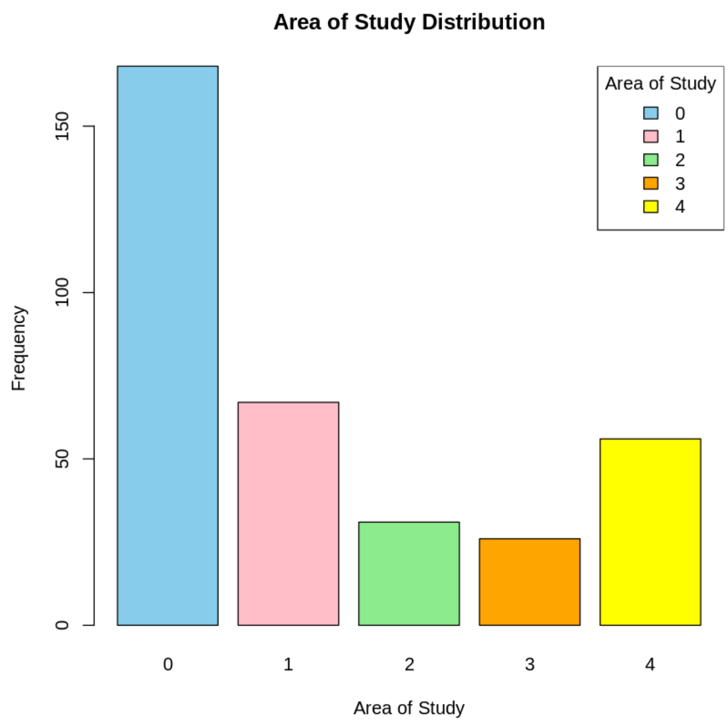
```
> aos_freq <- table(AOS2)
> print(aos_freq)

> aos_percent <- prop.table(aos_freq) * 100
> print (aos_freq)
```

	Science & Life Science	Engineering	Medical	commerce and management	arts and humanities
	168	67	31	26	56

Percentage	48.27	19.25	8.90	7.47	16.09

Barplot using R

```
> barplot(aos_freq, main = "Area of Study Distribution", xlab = "Area of Study", ylab = "Frequency", col = colors)
> legend("topright", legend = names(aos_freq), fill = colors, title = "Area of Study")
```



3.1.4 Total Screen Time

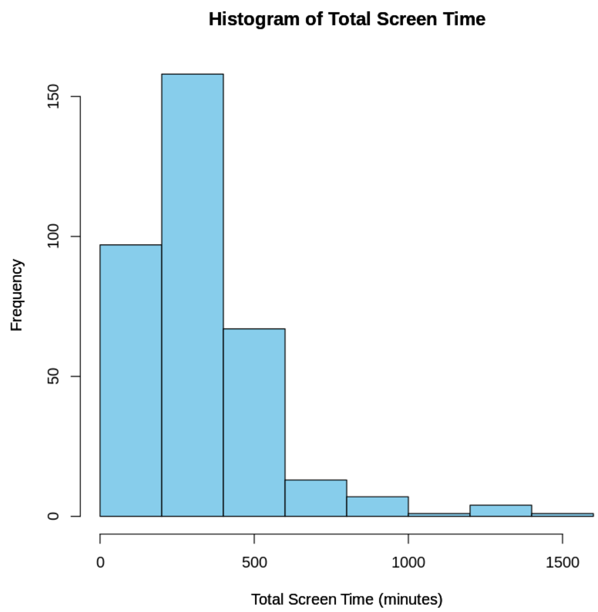
For the continuous variable Total Screen Time (in minutes), we can calculate the mean, median, standard deviation, minimum, and maximum values to describe the central tendency and dispersion of screen time in the sample.

```
>summary(TST)
```

Mean Of Total Screen Time (in minutes)	Median Total Screen Time (in minutes)	Standard Deviation of Total Screen Time (in minutes)	Minimum Screen Time (in minutes)	Maximum Screen Time (in minutes)
335.462643678161	300	217.15354343112	30.0	700

Histogram using R

```
> hist(TST, col = "skyblue", border = "black", main = "Histogram of Total Screen Time",
  xlab = "Total Screen Time (minutes)", ylab = "Frequency")

> title(main = "Histogram of Total Screen Time", xlab = "Total Screen Time (minutes)", ylab =
"Frequency")
```



3.1.5 Total Social Media Usage Time

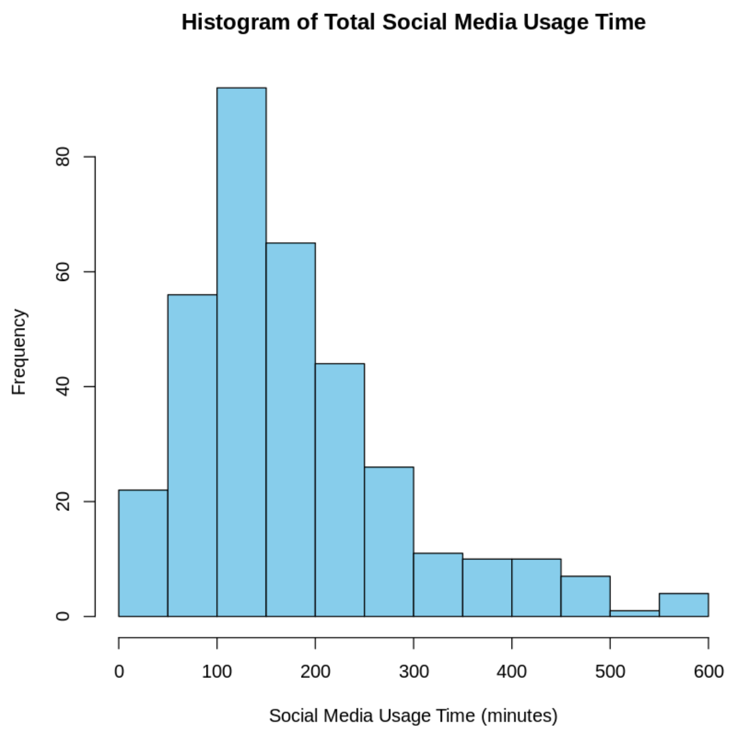
Similar to Total Screen Time, we can calculate the mean, median, standard deviation, minimum, and maximum values for Total Social Media Usage Time (in minutes) to describe the social media usage patterns in the sample.

>summary(TSTR)

Mean Total Social Media Usage Time (in minutes)	Median Total Social Media Usage Time (in minutes)	Standard Deviation of Total Social Media Usage Time (in minutes)	Minimum Total Social Media Usage Time (in minutes)	Maximum Total Social Media Usage Time (in minutes)
181.76724137931	160	113.560279243065	5.0	600

Histogram

> hist(TSTR, col = "skyblue", border = "black", main = "Histogram of Total Social Media Usage Time", xlab = "Social Media Usage Time (minutes)", ylab = "Frequency")



3.1.6 Stress Score, Depression Score, and Anxiety Score

These continuous variables representing mental health measures can also be described using the mean, median, standard deviation, minimum, and maximum values to understand the distribution of stress, depression, and anxiety in the sample.

> summary (STRESS)

> summary (DEPRESSION)

> summary (ANXIETY)

	Min.	1st Qu.	Median	Mean	3rd Qu	max

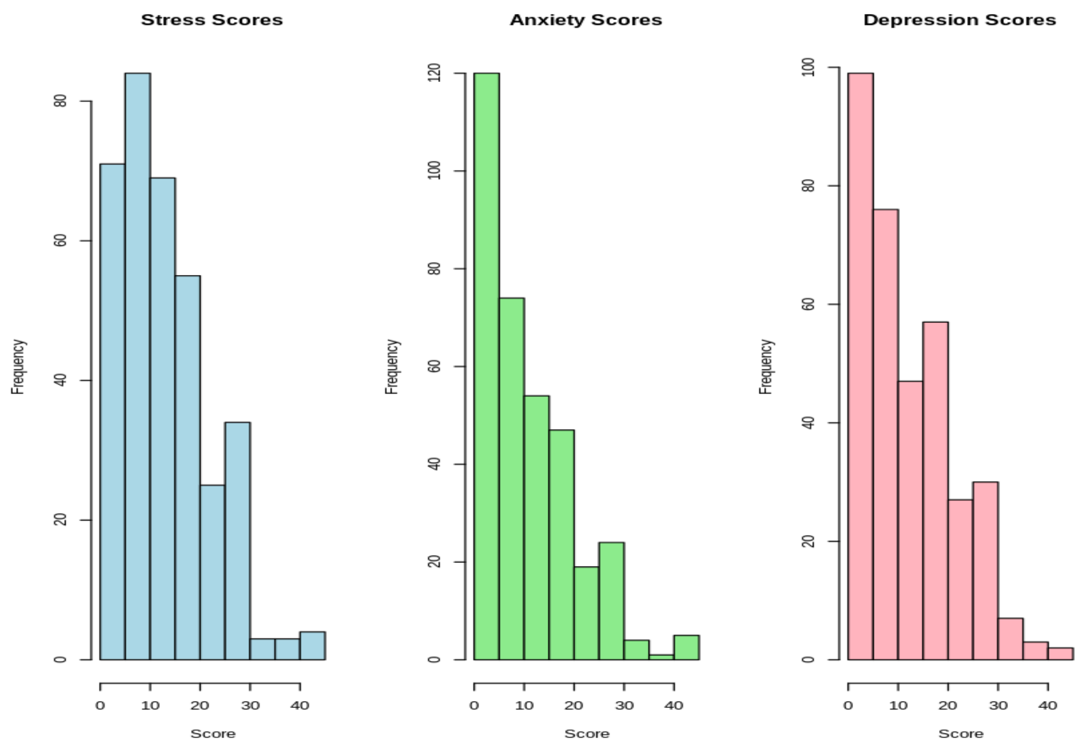
STRESS	0.0	6.0	12.0	13.03	18.0	42.0
DEPRESSION	0.0	4.0	10.0	12.3	18.00	42.0
ANXIETY	0.0	4.0	10.0	11.01	16.0	42.0

```
> par(mfrow=c(1,3))
```

> hist(STRESS, col = "lightblue", border = "black", main = "Stress Scores", xlab = "Score", ylab = "Frequency")

> hist(ANXIETY, col = "lightgreen", border = "black", main = "Anxiety Scores", xlab = "Score", ylab = "Frequency")

> hist(DEPRESSION, col = "lightpink", border = "black", main = "Depression Scores", xlab = "Score", ylab = "Frequency")



3.1.7 Cyberbullying Experience

The variable "Have You Ever Experienced Cyberbullying" has 3 categories: yes, no, and maybe. We can calculate the frequency and percentage of each category to understand the prevalence of cyberbullying experiences in the sample.

> CBull_freq <- table(CBull)

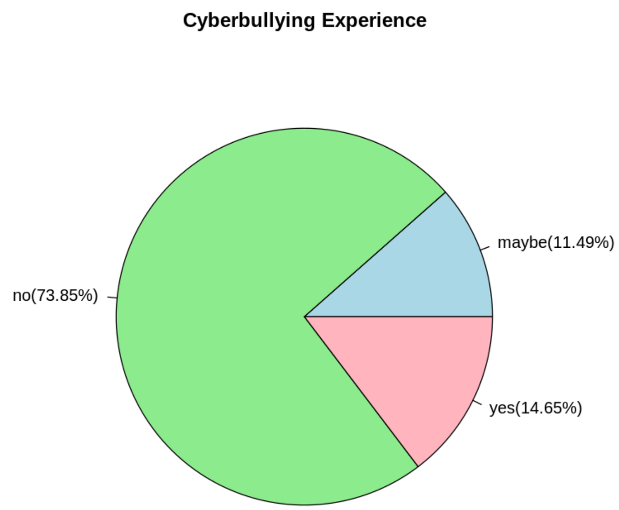
> print(CBull_freq)

> CBull_percent <- prop.table(CBull_freq) * 100

> print(CBull_percent)

	Maybe	No	yes
	40	257	51
Percentage	11.49	73.85	14.65

```
> pie(CBull_freq, labels = c("no", "maybe", "yes"), main = "Cyberbullying Experience",
   col = c("lightblue", "lightgreen", "lightpink"))
```



3.2 Correlation Analysis

Null Hypothesis (H0):

$\varrho TST, DEPR = \varrho TST, ANX = \varrho TST, STR = \varrho TSTR, DEPR = \varrho TSTR, ANX = \varrho TSTR, STR = 0$

Alternate Hypothesis (H1):

At least one correlation coefficient (ϱ) between social media usage and mental health measures among university students is not equal to zero.

3.2.1 Calculation of the Correlation Matrix:

Using an appropriate statistical software (e.g., R) to calculate the Pearson correlation coefficients between all pairs of the continuous variables.

The correlation matrix will show the strength and direction of the linear relationships between the variables.

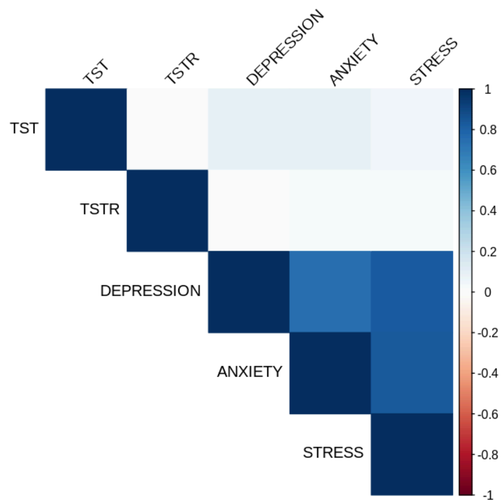
R command:

```
> mental_data<-data.frame(TST,TSTR, DEPRESSION, ANXIETY, STRESS)
> correlation_matrix <- cor(mental_data)
> correlation_matrix
```

	TST	TSTR	DEPR	ANX	STR
TST	1.0000000	0.01845588	0.01845588	0.099961	0.05517505
TSTR	0.01845588	1.0000000	0.0008057903	0.02955705	0.03322909
DEPR	0.0942862	0.0008057903	1.000000	0.75002570	0.82730878
ANX	0.09996199	0.029557	0.7500257012	0.7500257012	0.83317058
STR	0.05517505	0.0332290927	0.8273087754	0.83317058	1.0000000

3.2.2 Visualization of correlation matrix

```
# Visualize correlation matrix as a heatmap
> corrplot(correlation_matrix, method = "color", type = "upper", tl.col = "black", tl.srt = 45)
```



Based on the correlation matrix these results have been obtained:

Relationship	Correlation Coefficient (r)	Strength of Correlation
TST - TSTR	0.0185	Very Weak
TST - DEPR	0.0942	Very Weak
TST - ANX	0.1000	Very Weak
TST - STR	0.0552	Very Weak
TSTR - DEPR	0.0008	Very Weak
TSTR - ANX	0.0296	Very Weak
TSTR - STR	0.0332	Very Weak
DEPR - ANX	0.7500	Strong
DEPR - STR	0.8273	Very Strong
ANX - STR	0.8332	Very Strong

3.2.4 Calculation of Correlation Coefficients and p-values:

```
> p_values <- assess_correlation_significance(correlation_matrix)
> print(p_values)
```

	TST	TSTR	DEPR	ANX	STR
TST	NA	0.6032039	0.408899248	0.394944235	0.343671491
TSTR	0.6032039	NA	0.226501396	0.248041525	0.270903737
DEPR	0.4088992	0.2265014	NA	0.025429988	0.270903737
ANX	0.3949442	0.2480415	0.025429988	NA	0.270903737
STR	0.3436715	0.2709037	0.009433068	0.008506157	NA

Based on the provided p-values and the significance of correlations:

Relationship	P-Value	Significance
TST - TSTR	0.6032	Not Significant
TST - DEPR	0.4089	Not Significant
TST - ANX	0.3949	Not Significant
TST - STR	0.3437	Not Significant
TSTR - DEPR	0.2265	Not Significant
TSTR - ANX	0.2480	Not Significant
TSTR - STR	0.2709	Not Significant
DEPR - ANX	0.0254	Significant
DEPR - STR	0.0094	Significant
ANX - STR	0.0085	Significant

In summary, based on the provided p-values, there are statistically significant correlations between Depression Score, Anxiety Score, and Stress Score, while the correlations between Total Screen Time, Total Social Media Usage Time, and mental health measures are not statistically significant.

3.3 Group Comparisons Analysis

Gender:

Hypothesis: Gender may influence the levels of Total Screen Time, Total Social Media Usage Time, Stress Score, Depression Score, Anxiety Score, and the likelihood of experiencing cyberbullying.

Statistical Test: Conduct independent samples t-tests to compare the mean values of the continuous variables between male and female participants.

Welch Two Sample t-test

```
# Independent samples t-tests for each continuous variable with gender
> t_test_results <- list()
> variables <- c("TST", "TSTR", "STRESS", "DEPRESSION", "ANXIETY")
for (variable in variables) {
  t_test_results[[variable]] <- t.test(data[[variable]] ~ data$Gender)
}
```

$TST vs Gender

Welch Two Sample t-test

data: data[[variable]] by data$Gender

t-value	**df**	**p-value**
-2.7403	149.49	0.006887

alternative hypothesis: true difference in means between group female and group male is not equal to 0

95 percent confidence interval:	-122.26326	-19.81358

sample estimates:

mean in group female	mean in group male
281.5714	352.6098

$TSTR

Welch Two Sample t-test

data: data[[variable]] by data$Gender

t-value	**df**	**p-value**
-1.448	166.2	0.1495

alternative hypothesis: true difference in means between group female and group male is not equal to 0

95 percent confidence interval:	-44.189563	6.795624

sample estimates:

mean in group female	mean in group male
167.5833	186.2803

$STRESS

Welch Two Sample t-test

data: data[[variable]] by data$Gender

t-value	**df**	**p-value**
2.9824	141.83	0.003368

alternative hypothesis: true difference in means between group female and group male is not equal to 0

95 percent confidence interval:	1.138117	5.612965

sample estimates:

mean in group female	mean in group male
15.59524	12.21970

$DEPRESSION

Welch Two Sample t-test

data: data[[variable]] by data$Gender

t-value	df	p-value
2.1189	134.22	0.03594

alternative hypothesis: true difference in means between group female and group male is not equal to 0

95 percent confidence interval:	0.1742058	5.0573960

sample estimates:

mean in group female	mean in group male
14.21429	11.59848

$ANXIETY

Welch Two Sample t-test

data: data[[variable]] by data$Gender

t-value	**df**	**p-value**
3.1197	130.96	0.002227

alternative hypothesis: true difference in means between group female and group male is not equal to 0

95 percent confidence interval:	1.386734	6.193353

sample estimates:

mean in group female	mean in group male
13.88095	10.09091

I. **Total Screen Time (TST):**

- The t-test yielded:

t value	P value
-2.7403	0.006887.

- The p-value indicates that there is a statistically significant difference in the mean total screen time between male and female participants.

The mean total screen time for female participants(minutes)	while for male participants(minutes)
281.5714	352.6098.

II. <u>**Total Social Media Usage Time (TSTR):**</u>

- The t-test yielded:

t-value	**p-value**
-1.448	0.1495.

- The p-value indicates that there is no statistically significant difference in the mean total social media usage time between male and female participants.

The mean total screen time for female participants(minutes)	while for male participants(minutes)
167.5833	186.2803

III. <u>Stress Score:</u>

- The t-test yielded:

t-value	p-value
2.9824	0.003368.

- The p-value indicates that there is a statistically significant difference in the mean stress score between male and female participants.

The mean total screen time for female participants(minutes)	while for male participants(minutes)
15.59524	12.21970.

IV. <u>Depression Score:</u>

- The t-test yielded

t-value	p-value
2.1189	- 0.03594.

- The p-value indicates that there is a statistically significant difference in the mean depression score between male and female participants.

The mean total screen time for female participants(minutes)	while for male participants(minutes)
14.21429	11.59848

V. **Anxiety Score:**

- The t-test yielded:

t-value	p-value
3.1197	0.002227

- The p-value indicates that there is a statistically significant difference in the mean anxiety score between male and female participants.

The mean total screen time for female participants(minutes)	while for male participants(minutes)
13.88095	10.09091.

VI. **Likelihood of Experiencing Cyberbullying:**

- The chi-squared test yielded:

t-value	p-value
0.46643	0.792.

- The p-value indicates that there is no statistically significant association between the likelihood of experiencing cyberbullying and gender.

The t-test is not applicable due to the non-normal distribution of the continuous variables. Therefore, the Mann-Whitney U test, a non-parametric alternative, has been applied to analyze the data.

3.4 Mann Whittney U test

The Mann-Whitney U test, also known as the Wilcoxon rank-sum test, is a nonparametric statistical test used to compare two independent samples to determine whether they come from the same distribution. Unlike parametric tests such as the t-test, it does not assume that the data are normally distributed, making it particularly useful when dealing with small sample sizes or ordinal data. The test works by ranking all the values from both samples together and then examining the ranks of the values within each sample. The test statistic, U, is calculated based on the sum of the ranks for each sample. The null hypothesis of the Mann-Whitney U test is that the distributions of the two populations are equal, meaning that the likelihood of observing a particular value is the same in both samples. If the calculated U statistic is significantly different from what would be expected under the null hypothesis, the null hypothesis can be rejected, suggesting a significant difference between the two samples. The Mann-Whitney U test is widely used in fields such as medicine, psychology, and social sciences due to its robustness and flexibility in handling non-normally distributed data.

R code

```
library
library(stats)

> wilcox.test(TST ~ Gender, data = data)
```

```
> wilcox.test(TSTR ~ Gender, data = data)

> wilcox.test(STRESS ~ Gender, data = data)

> wilcox.test(Depression_Score ~ Gender, data = data)

> wilcox.test(Anxiety_Score ~ Gender, data = data)
```

Result

	P-value	**W**	**Hypothesis**
TST	0.00105	8457	Alternative(H1): The probability distribution of the two groups are not equal
TSTR	0.4194	10440	Alternative(H1): The probability distribution of the two groups are not equal
DEPRESSION	0.02615	12869	Alternative(H1): The probability distribution of the two groups are not equal
ANXIETY	0.001001	13721	Alternative(H1): The probability

			distribution of the two groups are not equal
STRESS	0.003322	13439	Alternative(H1): The probability distribution of the two groups are not equal

The results suggest that gender may influence stress, depression, and anxiety scores, as well as total screen time, but not total social media usage time or the likelihood of experiencing cyberbullying. “The results suggest males experience more stress, depression and anxiety as compared to females”.

3.5 Chi squared test between categorical variables

1. Gender vs CBull
2. Gender vs AOS2
3. Gender vs LOS
4. AOS2 vs LOS
5. AOS2 vs CBull
6. CBull vs LOS

#H0: there is not any significant association between two categorical variables
#H1: the two attributes are associates to each other significantly

(1). contingency_table1 <- table(data$Gender, data$CBull)
contingency_table1

	maybe	no	yes

female	8	64	12
male	32	193	39

chisq_test1<-chisq.test(contingency_table1)

chisq_test1

Pearson's Chi-squared test

data: contingency_table1

X-squared = 0.46643, df = 2, p-value = 0.792

Conclusion: no association between the variables

(2) contingency_table2 <- table(data$Gender, data$AOS2)

contingency_table2

#H0: there is not any significant association between two categorical variables

#H1: the two attributes are associates to each other significantly

chisq_test2<-chisq.test(contingency_table2)

chisq_test2

	0	1	2	3	4
female	32	17	7	10	18
male	136	50	24	16	38

Pearson's Chi-squared test

data: contingency_table2

X-squared = 7.3469, df = 4, p-value = 0.1187

Conclusion: no association between the variables

(3)contingency_table3 <- table(data$Gender, data$LOS)

contingency_table3

chisq_test3<-chisq.test(contingency_table3)

chisq_test3

	P.HD	PG	UG
female	3	10	71
male	17	44	203

Pearson's Chi-squared test

data: contingency_table3

X-squared = 2.3144, df = 2, p-value = 0.314

Conclusion: no association between the variables

(4) contingency_table4 <- table(data$LOS, data$AOS2)

contingency_table4

chisq_test4<-chisq.test(contingency_table4)

chisq_test4

	0	1	2	3	4
P.HD	12	1	2	3	2
PG	28	8	3	2	13
UG	128	58	26	21	41

Pearson's Chi-squared test

data: contingency_table4

X-squared = 10.22, df = 8, p-value = 0.2499

Conclusion: no association between the variables

(5) contingency_table5 <- table(data$AOS2, data$CBull)

contingency_table5

chisq_test5<-chisq.test(contingency_table5)

chisq_test5

	maybe	no	yes
0	20	117	31
1	7	51	9
2	6	21	4
3	3	22	1
4	4	46	6

Pearson's Chi-squared test

data: contingency_table5

X-squared = 8.6653, df = 8, p-value = 0.3713

Conclusion: no association between the variables

(6) contingency_table6<- table(data$LOS, data$CBull)

contingency_table6

chisq_test6<-chisq.test(contingency_table5)

chisq_test6

	maybe	no	yes

P.HD	5	11	4
PG	5	42	7
UG	30	204	40

Pearson's Chi-squared test

data: contingency_table5

X-squared = 8.6653, df = 8, p-value = 0.3713

Conclusion: no association between the variables

3.4 Shapiro Wilk Test for Normality

The Shapiro-Wilk test is a widely utilized statistical test for assessing the normality of a dataset. It evaluates whether a given sample comes from a normally distributed population. This test is particularly important in validating the assumptions of many parametric tests, which require normally distributed data for accurate results. The Shapiro-Wilk test operates by comparing the order statistics of the sample to the expected values under a normal distribution. Specifically, it calculates a W statistic, which measures how well the sample data conforms to a normal distribution. A W value close to 1 indicates that the data is consistent with a normal distribution, while a lower W value suggests deviations from normality.

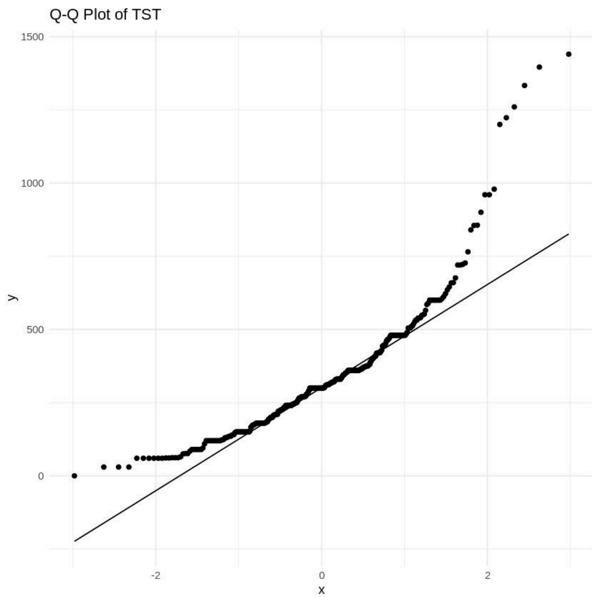
The null hypothesis of the Shapiro-Wilk test is that the sample data follows a normal distribution. If the p-value obtained from the test is below a predefined significance level (commonly 0.05), the null hypothesis is rejected, indicating that the data does not follow a normal distribution. The Shapiro-Wilk test is preferred over other normality tests due to its high power, particularly with small sample sizes, making it a crucial tool in preliminary data analysis and ensuring the validity of subsequent statistical procedures in fields such as biology, finance, and engineering.

Hypothesis setup

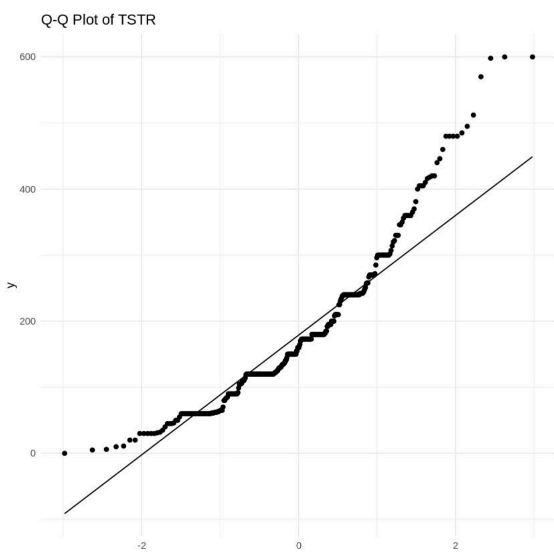
H0: The data follows a normal distribution

H1: The data does not follow a normal distribution

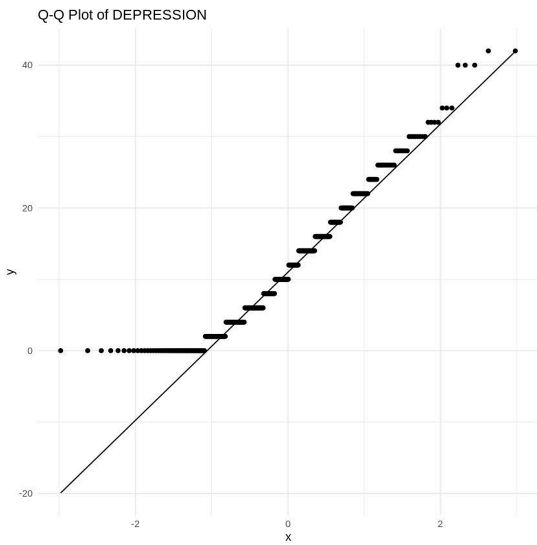
> shapiro.test(mental_data$TST)



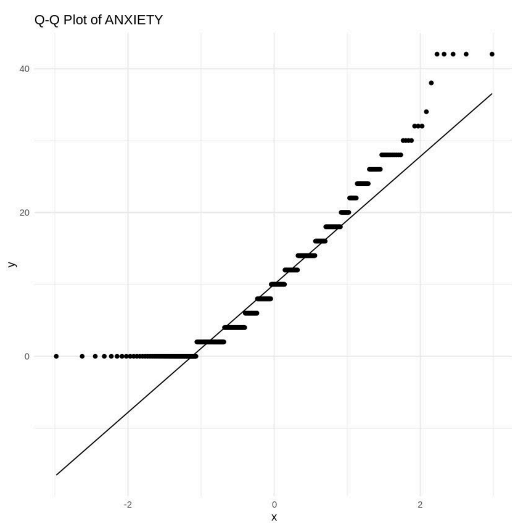
> shapiro.test(mental_data$TSTR)



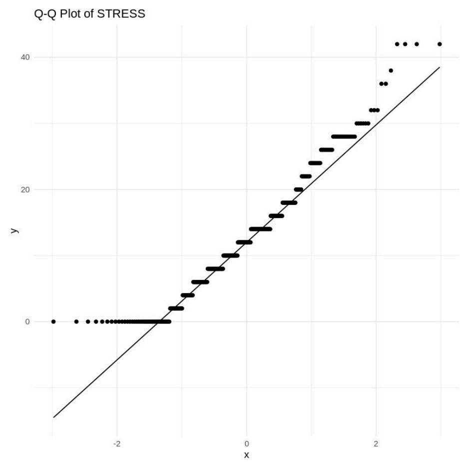
> shapiro.test(mental_data$DEPRESSION)



> shapiro.test(mental_data$ANXIETY)



> shapiro.test(mental_data$STRESS)



Results

	p-value	W
TST	<2.2e-16	0.85222
TSTR	4.909e-13	0.91593
DEPRESSION	9.345e-11	0.93908
ANXIETY	4.019e-13	0.91495
STRESS	4.4e-09	0.95312

The results suggest that the continuous variables aren't following normal distribution.

3.5 Regression Analysis

Regression is a statistical method used to establish a relationship between a dependent variable and one or more independent variables. In simpler terms, regression analysis helps to find the relationship between a dependent variable (outcome) and one or more independent variables (predictors). Regression analysis can be used to explore and understand the relationships between variables and to predict the future values of the dependent variable based on the values of the independent variable(s). Regression analysis is often used in many fields, including economics, finance, marketing, biology, psychology, and engineering. It is used to investigate the causal relationship between variables and make predictions based on historical data. There are many types of regression models, including linear regression, logistic regression, polynomial regression, and multiple regression. The choice of model depends on the nature of the data and the research question being addressed.

Linear regression is one of the most widely used regression models. It assumes a linear relationship between the dependent variable and the independent variable(s). The goal of linear regression is to find the best-fit line that explains the relationship between the variables. Multiple regression is used when there are two or more independent variables that affect the dependent variable. The goal of multiple regression is to find the best-fit line that explains the relationship between the variables. Regression analysis is a powerful tool for understanding and predicting the relationships between variables.

A linear model specifies a linear relationship between a dependent variable and independent variables.

$$y = \beta 0 + \beta 1 * X1 + \ldots\ldots\ldots\ldots\ldots + \beta n * Xn + \in$$

where y is the dependent variable, {Xi} are independent variables, {βi} are parameters of the model.

Assumptions of Regression Analysis

- Linearity: The relationship between the dependent variable and the independent variable(s) should be linear.
- Independence: The observations should be independent of each other.
- Homoscedasticity: The variance of the errors should be constant across the range of the independent variable(s).
- Normality: The errors should be normally distributed.
- No multicollinearity: The independent variables should not be highly correlated with each other.

Hypothesis setup

H0: data is normally distributed

H1: data does not follow normality

To perform regression analysis, the assumption of normality for the residuals is required. The Shapiro-Wilk test is commonly used to assess if the data follows a normal distribution.

If the Shapiro-Wilk test indicates that the data does not meet the normality assumption (p-value < 0.05), the Box-Cox power transformation can be applied to normalize the data. This transformation introduces a parameter λ that is estimated to optimize the normality of the transformed data.

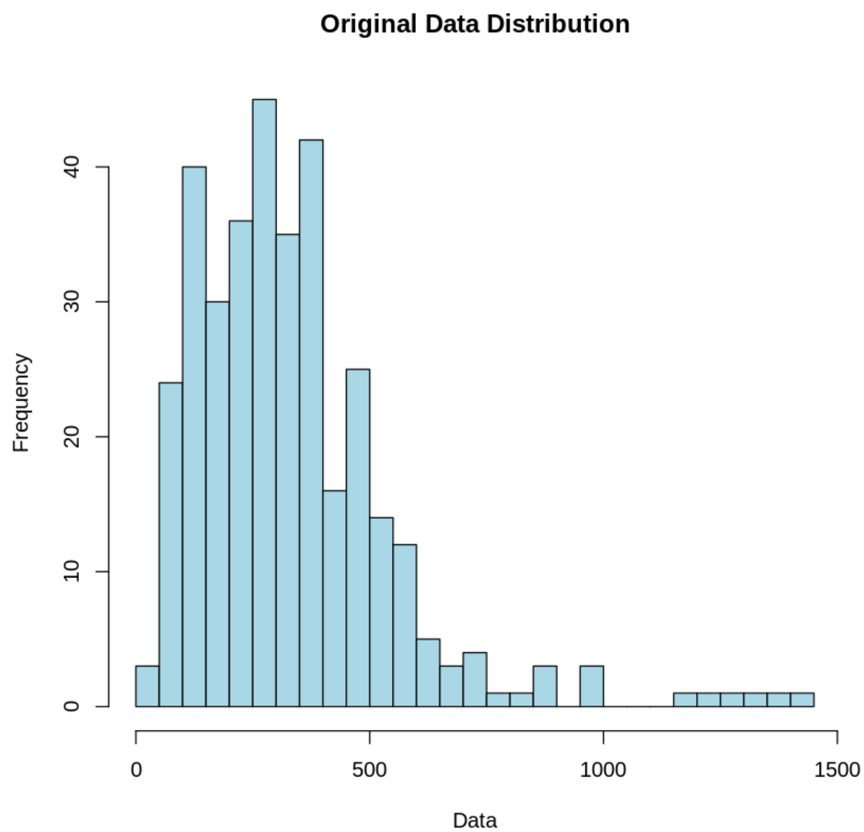
3.5.1 Box-Cox Transformation

The Box-Cox transformation is a method used to transform non-normal data into a normal distribution. The transformation involves raising the data to the power of lambda (λ), which can vary from -5 to 5. The formula for the Box-Cox transformation is:

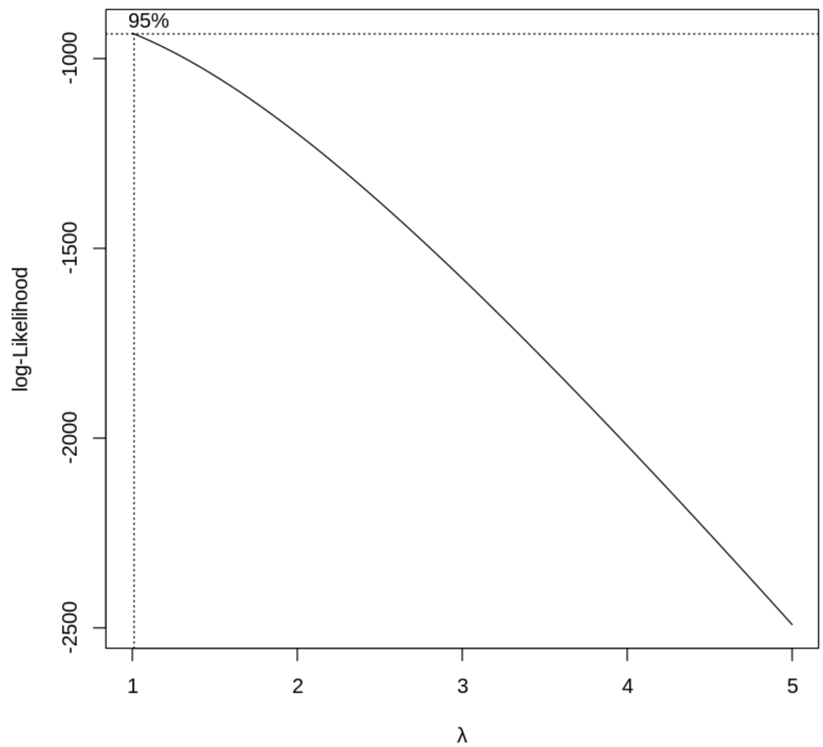
$$y(\lambda) = \begin{cases} \frac{y^\lambda - 1}{\lambda}, & \text{if } \lambda \neq 0 \\ \log(y), & \text{if } \lambda = 0 \end{cases}$$

i. Box-Cox Transformation of Total Screen Time

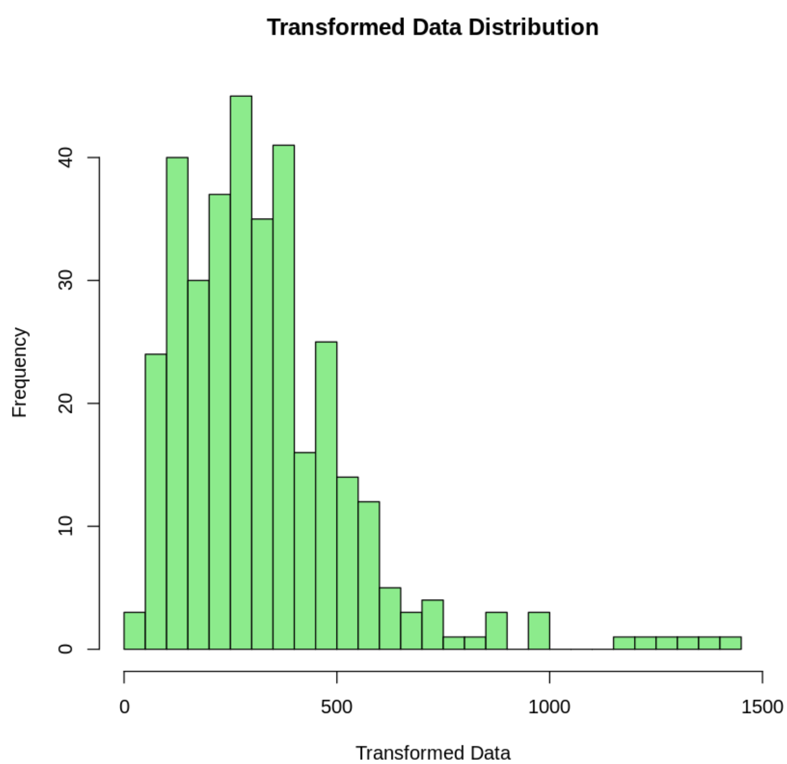
> hist(TST, breaks = 30, main = "Original Data Distribution", xlab = "Data", col = "lightblue", border = "black")




```
> bc <- boxcox(TST ~ 1, lambda = seq(1, 5, by = 0.1)) # Using a range of lambda values
> best_lambda <- bc$x[which.max(bc$y)] # Extracting the best lambda
> transformed_data <- (TST^best_lambda - 1) / best_lambda
> hist(transformed_data, breaks = 30, main = "Transformed Data Distribution", xlab =
"Transformed Data", col = "lightgreen", border = "black")
```



95%
-1000
-1500
-2000
-2500
log-Likelihood
1
2
3
4
5
λ

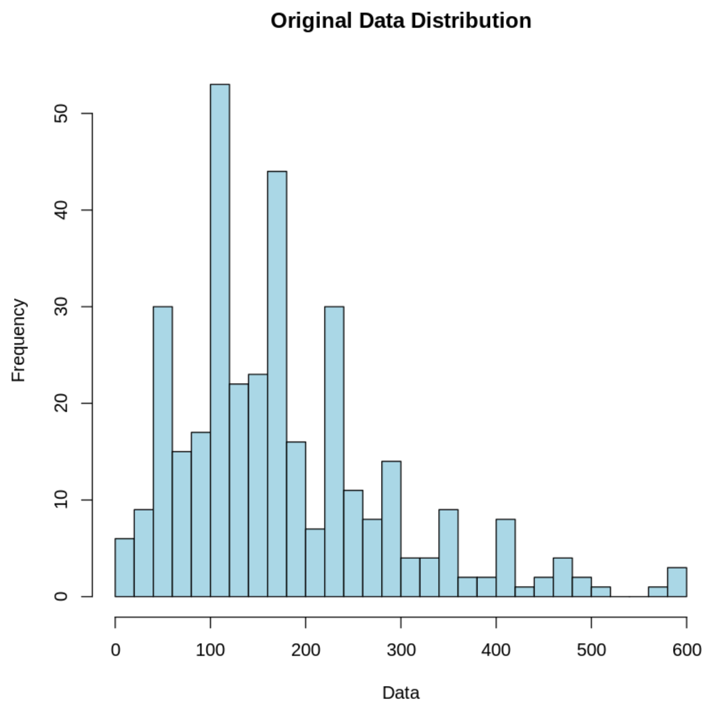


Transformed Data Distribution
Frequency
0
10
20
30
40
0
500
1000
1500
Transformed Data

ii. Box-Cox Transformation of TSTR

> primary_data2 <- c(TSTR)

hist(primary_data2, breaks = 30, main = "Original Data Distribution", xlab = "Data", col = "lightblue", border = "black")

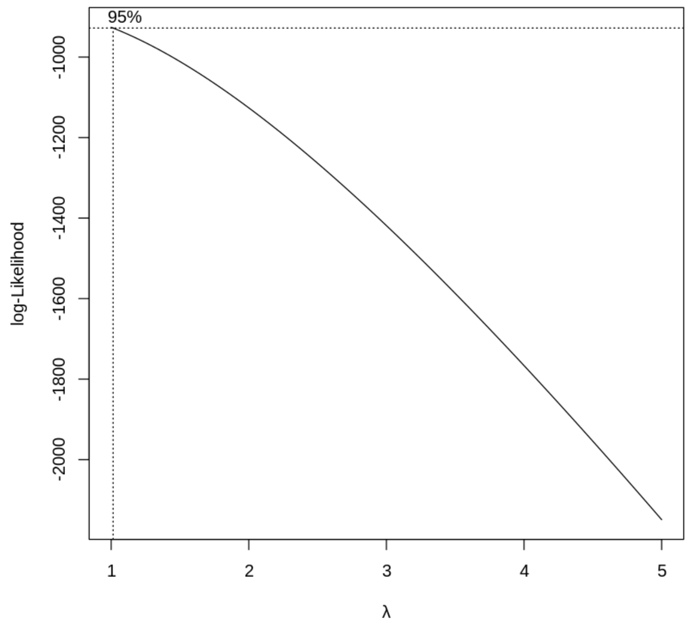


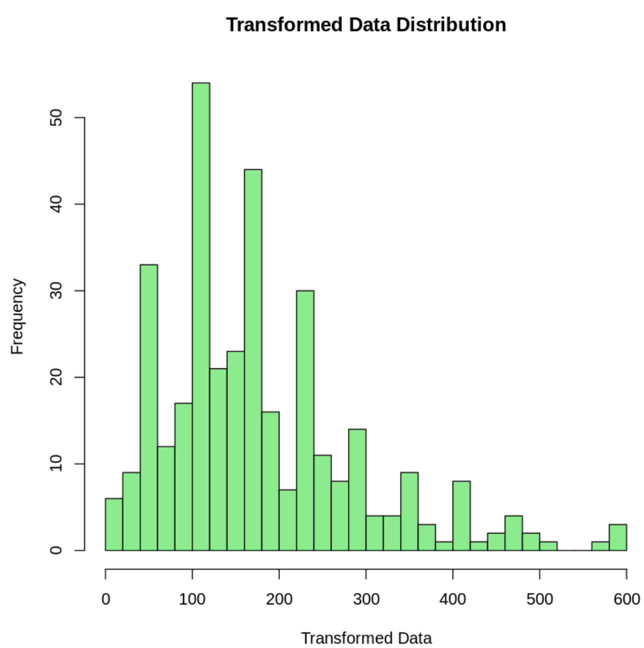
> bc <- boxcox(primary_data2 ~ 1, lambda = seq(1, 5, by = 0.1)) # Using a range of lambda values

> best_lambda <- bc$x[which.max(bc$y)] # Extracting the best lambda

> transformed_data <- (primary_data2^best_lambda - 1) / best_lambda

> hist(transformed_data, breaks = 30, main = "Transformed Data Distribution", xlab = "Transformed Data", col = "lightgreen", border = "black")





3.5.2 Linear Regression

Linear regression is a statistical method used to model the linear relationship between a dependent variable and one or more independent variables. It fits a linear equation to the observed data, allowing for predictions and understanding the influence of each independent variable on the dependent variable.

The formula for a simple linear regression (one independent variable) is:

$$y = \beta_0 + \beta_1 x + \epsilon$$

i. Depression as depended variable

> DEPR_model<-lm(DEPRESSION ~ Gender+AOS2+LOS+STST+TSTR,data=data)

summary(DEPR_model)

Call:

lm(formula = DEPRESSION ~ Gender + AOS2 + LOS + TST + TSTR, data = data)

Results

Residuals

Min	1Q	Median	3Q	Max
-16.2837	-8.0224	-0.5729	6.4757	30.9517

Summary model:

Coefficients:	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	12.6888955	2.6519199	4.785	2.55e-06 ***
Gender male	-2.9959968	1.2326905	-2.430	0.0156 *
AOS2	-0.1333247	0.3456371	-0.386	0.6999

LOSPG	-0.7282720	2.5269646	-0.288	0.7734
LOSUG	0.3065026	2.2364813	0.137	0.8911
STST	0.0050785	0.0024116	2.106	0.0359 *
TSTR	0.0008006	0.0045586	0.176	0.8607

---Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 9.603 on 341 degrees of freedom

Multiple R-squared: 0.02823, Adjusted R-squared: 0.01113

F-statistic: 1.651 on 6 and 341 DF, p-value: 0.1324

ii. Anxiety score as depended variable

> ANX_model<-lm(ANXIETY ~ Gender+AOS2+LOS+STST+TSTR,data=data) summary(ANX_model)

lm(formula = ANXIETY ~ Gender + AOS2 + LOS + STST + TSTR, data = data)

Residuals:

Min	1Q	Median	3Q	Max
-14.798	-7.129	-1.307	5.411	33.700

Coefficients:	Estimate	Std. Error	T value	Pr(>|t|)	

(Intercept)	12.980434	2.559146	5.072	6.47e-07	***
Gender male	-4.312033	1.189567	-3.625	0.000333	***
AOS2	-0.201519	0.333546	-0.604	0.546130	
LOSPG	-1.848805	2.438563	-0.758	0.448883	
LOSUG	-0.823588	2.158241	-0.382	0.702995	
STST	0.005432	0.002327	2.334	0.020166	*
TSTR	0.003618	0.004399	0.822	0.411436	

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 9.267 on 341 degrees of freedom

Multiple R-squared: 0.05037, Adjusted R-squared: 0.03366

F-statistic: 3.015 on 6 and 341 DF, p-value: 0.006949

iii Stress Scores as depended variable

```
> STR_model<-lm(STRESS ~ Gender+AOS2+LOS+STST+TSTR+CBull,data=data)
summary(STR_model)
lm(formula = STRESS ~ Gender + AOS2 + LOS + STST + TSTR + CBull,
data = data)
Residuals:
```

Min	1Q	Median	3Q	Max
-17.6205	-6.9938	-0.9413	5.7628	30.5100

Coefficients:	Estimate	Std.Error	tvalue	Pr(>|t|)	
(Intercept)	13.710668	2.822925	4.857	1.82e-06	***
Gendermale	-3.633655	1.172911	-3.098	0.00211	**
AOS2	0.081413	0.330897	0.246	0.80580	
LOSPG	0.173796	2.419068	0.072	0.94277	
LOSPG	0.173796	2.419068	0.072	0.94277	
LOSUG	0.718413	2.140523	0.336	0.73736	
STST	0.003366	0.002299	1.464	0.14403	
TSTR	0.003698	0.004343	0.851	0.39510	
CBullno	-0.829245	1.569215	-0.528	0.59754	
CBullyes	1.344425	1.939785	0.693	0.48873	

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 9.136 on 339 degrees of freedom

Multiple R-squared: 0.04031, Adjusted R-squared: 0.01767

F-statistic: 1.78 on 8 and 339 DF, p-value: 0.0799

iv. Depression vs other mental health score

> among_model<-lm(DEPRESSION ~ ANXIETY+STRESS,data=data)

summary(among_model)

Call:

lm(formula = DEPRESSION ~ ANXIETY + STRESS, data = data)

Residuals:

Min	1Q	Median	3Q	Max
-15.6566	-2.7078	-0.5174	2.8667	17.9186

Coefficients:	Estimate	Std. Error	T value	Pr(>|t|)	
(Intercept)	0.95342	0.49584	1.923	0.055322	.
ANXIETY	0.20344	0.05494	3.703	0.000248	***
STRESS	0.69335	0.05619	12.340	< 2e-16	***

v. Anxiety vs other mental health score

>among_model1<-lm(ANXIETY ~DEPRESSION +STRESS,data=data)

> summary(among_model1)

lm(formula = ANXIETY ~ DEPRESSION + STRESS, data = data)

Residuals:

Min	1Q	Median	3Q	Max
-21.1560	-2.6502	0.1387	3.1249	15.4993

Coefficients:	Estimate	Std. Error	T value	Pr(>|t|)	
(Intercept)	-0.27578	0.47883	-0.576	0.565032	
DEPRESSIO N	0.18789	0.05074	3.703	0.000248	***
STRESS	0.68922	0.05316	12.966	<2e-16	***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5.127 on 345 degrees of freedom

Multiple R-squared: 0.7059, Adjusted R-squared: 0.7042

F-statistic: 414 on 2 and 345 DF, p-value: < 2.2e-15

vi. Stress vs others mental health score

> among_model12<-lm(STRESS ~DEPRESSION +ANXIETY,data=data)

summary(among_model12)

lm(formula = STRESS ~ DEPRESSION + ANXIETY, data = data)

Residuals:

Min	1Q	Median	3Q	Max
-13.2085	-2.7128	-0.2621	2.4267	14.9314

Coefficients:	Estimate	Std.Error	tvalue	Pr(>|t|)	
(Intercept)	2.40127	0.37626	6.382	5.64e-10	***
DEPRESSION	0.44166	0.03579	12.340	<2e-16	***
ANXIETY	0.47536	0.03666	12.966	<2e-16	***

---Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.258 on 345 degrees of freedom

Multiple R-squared: 0.7878, Adjusted R-squared: 0.7866

F-statistic: 640.5 on 2 and 345 DF, p-value: < 2.2e-16

Chapter IV
Conclusion

Chapter - 04

Conclusion

Overview of the Analysis

The analysis presented aimed to investigate the impact of social media on mental health and cyberbullying among college students. A comprehensive approach was adopted, encompassing descriptive statistics, correlation analysis, group comparisons, and inferential statistical tests including the Mann-Whitney U test and chi-squared test. The findings provide valuable insights into the relationships between social media usage, mental health metrics, and demographic variables such as gender and area of study.

Descriptive Statistics

Descriptive statistics were utilized to summarize the key characteristics of the data, providing an initial understanding of the sample's composition and the distribution of variables.

Gender Distribution

The dataset comprised 348 participants, with a significant majority being male (264, 75.86%) and a smaller proportion female (84, 23.13%). This distribution is crucial for understanding the potential gender biases in subsequent analyses .

Level of Study

Participants were categorized into three levels of study: undergraduate (UG), postgraduate (PG), and PhD. The majority were undergraduates (274, 78.73%), followed by postgraduates (54, 15.51%), and PhD students (20, 5.74%) .

Area of Study

The area of study was divided into five categories: Science & Life Science, Engineering, Medical, Commerce & Management, and Arts & Humanities. The largest group was Science & Life Science (168, 48.27%), followed by Engineering (67, 19.25%), Medical (31, 8.90%), Commerce & Management (26, 7.47%), and Arts & Humanities (56, 16.09%) .

Screen Time and Social Media Usage

The mean total screen time was 335.46 minutes per day, with a standard deviation of 217.15 minutes, indicating substantial variability in screen time among participants. Similarly, the mean total social media usage time was 181.77 minutes per day, with a standard deviation of 113.56 minutes .

Mental Health Scores

The mental health of participants was assessed using the DASS-21 scores for stress, depression, and anxiety. The results showed considerable variability, with mean scores of 13.88 for stress, 14.21 for depression, and 13.88 for anxiety, reflecting the diversity in mental health status among the students .

Correlation Analysis

Correlation analysis was conducted to examine the relationships between screen time, social media usage, and mental health scores.

Key Findings

The analysis revealed significant correlations between depression, anxiety, and stress scores, indicating that these mental health issues are interrelated. However, the correlations between screen time, social media usage, and mental health measures were not statistically significant, suggesting that the amount of time spent on screens or social media alone does not directly correlate with mental health outcomes .

Group Comparisons Analysis

Group comparisons were performed to explore differences in screen time, social media usage, and mental health scores across gender and other demographic variables.

Gender Differences

Screen Time

A significant difference was found in total screen time between male and female participants, with males reporting higher screen time on average. This difference was statistically significant ($p = 0.006887$), indicating a gender disparity in screen usage habits .

Social Media Usage

No significant difference was observed in social media usage time between genders ($p = 0.1495$), suggesting that both male and female students spend similar amounts of time on social media .

Stress, Depression, and Anxiety

Significant differences were found in stress ($p = 0.003368$), depression ($p = 0.03594$), and anxiety scores ($p = 0.002227$) between genders. Females reported higher levels of stress, depression, and anxiety compared to males, highlighting gender-based differences in mental health .

Inferential Statistical Tests

Mann-Whitney U Test

The Mann-Whitney U test was applied to compare the distributions of continuous variables across gender due to the non-normal distribution of the data.

Key Results

The results corroborated the findings from the t-tests, showing significant differences in total screen time ($p = 0.00105$), depression ($p = 0.02615$), anxiety ($p = 0.001001$), and stress scores ($p = 0.003322$) between genders. These results emphasize the influence of gender on screen time and mental health outcomes .

Chi-Squared Test

The chi-squared test was used to examine associations between categorical variables, such as gender and cyberbullying experience.

Key Findings

No significant association was found between gender and the likelihood of experiencing cyberbullying ($p = 0.792$), suggesting that the experience of cyberbullying is not significantly influenced by gender .

Implications and Recommendations

Implications for Mental Health Interventions

The findings highlight the need for targeted mental health interventions that consider gender differences. Female students, in particular, may benefit from increased mental health support given their higher levels of reported stress, depression, and anxiety.

Addressing Screen Time Disparities

Given the significant difference in screen time between male and female students, interventions aimed at promoting balanced screen usage should be gender-sensitive, addressing the specific needs and habits of each group.

Comprehensive Cyberbullying Prevention

Although no significant gender difference was found in cyberbullying experiences, the overall impact of cyberbullying on mental health necessitates comprehensive prevention programs. These programs should include education on safe online behavior and support systems for victims of cyberbullying.

Future Research Directions

Further research should explore the underlying factors contributing to the gender differences in mental health outcomes and screen time. Longitudinal studies could provide deeper insights into how these variables interact over time. Additionally, examining other demographic variables, such as socio-economic status and academic performance, could offer a more nuanced understanding of the factors influencing mental health and social media usage.

Conclusion

This analysis underscores the complex relationships between social media usage, mental health, and demographic variables among college students. While screen time and social media usage are pervasive in modern student life, their direct impact on mental health appears to be mediated by other factors, including gender. By acknowledging and addressing these nuances, stakeholders can develop more effective strategies to support the mental health and well-being of college students in the digital age.

References

1 https://colab.research.google.com/

2 https://cran.r-project.org/

3 https://en.m.wikipedia.org/wiki/Digital_media_use_and_mental_health

4 https://en.m.wikipedia.org/wiki/Shapiro%E2%80%93Wilk_test

5 https://en.m.wikipedia.org/wiki/Linear_regression

6 https://novopsych.com.au/assessments/depression/depression-anxiety-stress-scales-short-form-dass-21/

7 https://www.simplilearn.com/tutorials/statistics-tutorial/chi-square-test

8 https://www.biochemia-medica.com/en/journal/20/1/10.11613/BM.2010.004/fullArticle

9 https://byjus.com/maths/non-parametric-test/

10 https://forms.gle/cutDf8xHUmb56L6AA (Google Form)

Appendix

DASS21

Name: Date:

Please read each statement and circle a number 0, 1, 2 or 3 which indicates how much the statement applied to you **over the past week**. There are no right or wrong answers. Do not spend too much time on any statement.

The rating scale is as follows:

0 Did not apply to me at all
1 Applied to me to some degree, or some of the time
2 Applied to me to a considerable degree or a good part of time
3 Applied to me very much or most of the time

1 (s)	I found it hard to wind down	0	1	2	3
2 (a)	I was aware of dryness of my mouth	0	1	2	3
3 (d)	I couldn't seem to experience any positive feeling at all	0	1	2	3
4 (a)	I experienced breathing difficulty (e.g. excessively rapid breathing, breathlessness in the absence of physical exertion)	0	1	2	3
5 (d)	I found it difficult to work up the initiative to do things	0	1	2	3
6 (s)	I tended to over-react to situations	0	1	2	3
7 (a)	I experienced trembling (e.g. in the hands)	0	1	2	3
8 (s)	I felt that I was using a lot of nervous energy	0	1	2	3
9 (a)	I was worried about situations in which I might panic and make a fool of myself	0	1	2	3
10 (d)	I felt that I had nothing to look forward to	0	1	2	3
11 (s)	I found myself getting agitated	0	1	2	3
12 (s)	I found it difficult to relax	0	1	2	3
13 (d)	I felt down-hearted and blue	0	1	2	3
14 (s)	I was intolerant of anything that kept me from getting on with what I was doing	0	1	2	3
15 (a)	I felt I was close to panic	0	1	2	3
16 (d)	I was unable to become enthusiastic about anything	0	1	2	3
17 (d)	I felt I wasn't worth much as a person	0	1	2	3
18 (s)	I felt that I was rather touchy	0	1	2	3
19 (a)	I was aware of the action of my heart in the absence of physical exertion (e.g. sense of heart rate increase, heart missing a beat)	0	1	2	3
20 (a)	I felt scared without any good reason	0	1	2	3
21 (d)	I felt that life was meaningless	0	1	2	3

DASS-21 Scoring Instructions

The DASS-21 should not be used to replace a face to face clinical interview. If you are experiencing significant emotional difficulties you should contact your GP for a referral to a qualified professional.

Depression, Anxiety and Stress Scale - 21 Items (DASS-21)

The Depression, Anxiety and Stress Scale - 21 Items (DASS-21) is a set of three self-report scales designed to measure the emotional states of depression, anxiety and stress.

Each of the three DASS-21 scales contains 7 items, divided into subscales with similar content. The depression scale assesses dysphoria, hopelessness, devaluation of life, self-deprecation, lack of interest / involvement, anhedonia and inertia. The anxiety scale assesses autonomic arousal, skeletal muscle effects, situational anxiety, and subjective experience of anxious affect. The stress scale is sensitive to levels of chronic non-specific arousal. It assesses difficulty relaxing, nervous arousal, and being easily upset / agitated, irritable / over-reactive and impatient. Scores for depression, anxiety and stress are calculated by summing the scores for the relevant items.

The DASS-21 is based on a dimensional rather than a categorical conception of psychological disorder. The assumption on which the DASS-21 development was based (and which was confirmed by the research data) is that the differences between the depression, anxiety and the stress experienced by normal subjects and clinical populations are essentially differences of degree. The DASS-21 therefore has no direct implications for the allocation of patients to discrete diagnostic categories postulated in classificatory systems such as the DSM and ICD.

Recommended cut-off scores for conventional severity labels (normal, moderate, severe) are as follows:

<u>NB</u> Scores on the DASS-21 will need to be multiplied by 2 to calculate the final score.

	Depression	**Anxiety**	**Stress**
Normal	0-9	0-7	0-14
Mild	10-13	8-9	15-18
Moderate	14-20	10-14	19-25
Severe	21-27	15-19	26-33
Extremely Severe	28+	20+	34+

Lovibond, S.H. & Lovibond, P.F. (1995). Manual for the Depression Anxiety & Stress Scales. (2nd Ed.)Sydney: Psychology Foundation.