King Saud University
College of Computer and Information Sciences
Information Technology Department

جامعة الملك سعود
King Saud University
كلية علوم الحاسب والمعلومات
قسم تقنية المعلومات

Logbook Entry:

Primary Data (Survey)

Group#1

Supervised By:

Dr. Khulood Alyahya

2024-2025

1446-1447

• Introduction

Dataset: The survey we conducted on the impact of social media on mental health.

Objective in EDA:

The objective of the Exploratory Data Analysis (EDA) is to systematically examine the survey data to uncover insights regarding the impact of social media on mental health. This involves visualizing distributions, identifying trends and correlations, and detecting any anomalies or patterns in the data that could inform further analysis and research on user experiences and emotional well-being in relation to social media usage

---

Logbook Entry: Exploratory Data Analysis (EDA)

Analysis Performed

- **Loading the Data and Reviewing a Sample of the Data**
  We loaded the data from a CSV file using the Pandas library in Python.

- Then**:**
  We extracted a random sample from the dataset to get a quick overview of its

contents and value distribution.

	the age:	Gender:	Area:	Current educational level:	marital status:	Employment status:	Do you use social media applications?	What social media platforms do you use?	What app do you use the most?	How many hours do you spend on social media platforms daily?	...	Do you feel that using social media has affected your ability to focus and accomplish daily tasks?	Do you think that consuming quick content (such as watching short videos and push notifications...) has affected your patience and ability to deal with long tasks?	Do you use social media right before going to sleep?	Do you have difficulty sleeping because of thinking about what you saw on social media platforms?	Does the number of likes or comments you get on your posts affect you?	Have you changed your opinion or feeling based on the reactions of others on social media platforms?	Do you prefer interacting with friends or family online rather than face-to-face?	How often do you find yourself using social media for longer than you planned?	How do you feel when you compare your life to the lives of others on social media?	What methods, if any, do you use to limit your social media access?
45	18-24	feminine	Riyadh	High school or equivalent	bachelor	student	Yes	Instagram, X (Twitter), LinkedIn, TikTok, Snap...	Youtube	6-May	...	Yes, a lot	Yes, a lot	Yes, always	sometimes	Yes, always	Yes, always	Yes, always	always	Dissatisfaction, search for idealism, self-exh...	Setting other goals that occupy time and benef...
673	18-24	male	Hail	Bachelor's degree	bachelor	student	Yes	Instagram, X (Twitter), TikTok, Snapchat, Yout...	Snapchat	8-Jul	...	Yes, a lot	No, never	Yes, always	rarely	I don't care about the number of likes or comm...	No, never	sometimes	sometimes	I don't feel it at all	Isolate the device
373	18-24	feminine	Sakaka	Bachelor's degree	bachelor	student	Yes	Instagram, Snapchat, Youtube, WhatsApp, Locket	Instagram	6-May	...	Yes, a lot	Yes, a lot	Yes, always	rarely	I don't care about the number of likes or comm...	rarely	No, never	always	Feeling like wasting time on what is useless	Sitting with the family
380	18-24	feminine	Al-Baha	High school or equivalent	bachelor	student	Yes	Facebook, Instagram, X (Twitter), LinkedIn, Tl...	Instagram	2-Jan	...	Yes, a lot	Yes, a lot	Yes, always	Yes, always	Yes, always	Yes, always	Yes, always	always	Frustration and why he has it and I don't	Leave the device
1	18-24	male	Riyadh	Bachelor's degree	bachelor	Not employed	Yes	Instagram, X (Twitter), LinkedIn, TikTok, Snap...	X (Twitter)	4-Mar	...	sometimes	rarely	sometimes	No, never	No, never	No, never	No, never	sometimes	I don't feel anything, thank God	Use the restrictions in Apple settings (from S...

5 rows × 22 columns

This step was crucial for understanding the overall structure of the data before diving into more in-depth exploratory analysis.

A. Structure Investigation

Through our exploration of the dataset, we gained a general understanding of its structure. By using tools like shape and dtypes, we observed that the dataset contains 851 rows and 22 columns. This gives us an initial idea of the The dataset consists of 22 columns, all of which are of the object data type. This means the variables are likely categorical or text-based, which will guide our next steps in terms of analysis, such as converting them to numerical values if needed or applying techniques that are suitable for categorical data. Understanding the data types helps us decide how to handle these features during preprocessing and analysis.

1.1. Structure of non-numerical features

Here, we will delve deeper and explore the non-numerical data in our dataset. Non-numerical features play an important role in understanding user behavior and their perspectives as well.

Sample of non-numerical data:

	Gender:	Area:	Current educational level:	marital status:	Employment status:	Do you use social media applications?	What social media platforms do you use?	What app do you use the most?	How do you feel when you compare your life to the lives of others on social media?	What methods, if any, do you use to limit your social media access?
0	Female	Riyadh	High school or equivalent	bachelor	student	Yes	Instagram, X (Twitter), TikTok, Snapchat, Yout...	TikTok	Others hate hatred	nothing
1	male	Riyadh	Bachelor's degree	bachelor	Not employed	Yes	Instagram, X (Twitter), LinkedIn, TikTok, Snap...	X (Twitter)	I don't feel anything, thank God	Use the restrictions in Apple settings (from S...
2	Female	Riyadh	Bachelor's degree	bachelor	student	Yes	Instagram, X (Twitter), LinkedIn, Snapchat, Yo...	WhatsApp	I don't feel anything	Setting limits on applications, placing the de...
3	Female	Riyadh	High school or equivalent	bachelor	student	Yes	X (Twitter), TikTok, Snapchat, Youtube, WhatsA...	WhatsApp	G	B
4	Female	Riyadh	Bachelor's degree	married	Housewife, unemployed	Yes	Instagram, X (Twitter), Snapchat, WhatsApp, Lo...	Snapchat	Against comparisons	Leave the device in the room

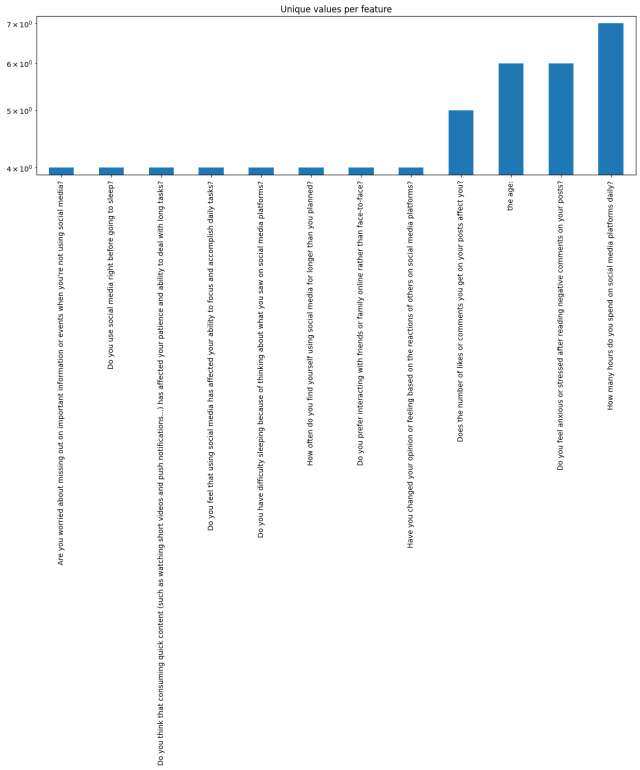
And here is a further breakdown of this data: the top, most frequent, and unique values. This chart helps us understand the data better by visualizing the distribution of these values and identifying patterns or trends in user behavior and perspectives.

	Gender:	Area:	Current educational level:	marital status:	Employment status:	Do you use social media applications?	What social media platforms do you use?	What app do you use the most?	How do you feel when you compare your life to the lives of others on social media?	What methods, if any, do you use to limit your social media access?
count	839	839	839	839	839	839	839	839	839	839
unique	2	20	7	4	17	1	245	14	547	607
top	Female	Riyadh	Bachelor's degree	bachelor	student	Yes	Instagram, X (Twitter), TikTok, Snapchat, Yout...	WhatsApp	I wish for a better life	nothing
freq	602	445	444	494	391	839	62	235	47	62

1.2. Structure of numerical features

Through the chart illustrating the number of unique values for each numerical feature, we observe that the feature "number of hours spent by users" contains a greater number of unique values than the feature "worried about not obtaining information." These features are arranged in ascending order, clearly showing the difference in the number of unique values between them.

This indicates that "number of hours spent" reflects greater diversity in the data compared to being worried, which may have a limited number of values. This suggests that user behavior regarding the amount of time spent on social media can vary significantly, while feelings of being worried may be more homogeneous.



1.3. Conclusion of structure investigation

At the conclusion of this first investigation, we now have a clearer understanding of the general structure of our dataset. Here's a summary of the key findings:

- Number of Samples (Objects): 851

- Number of Features (Attributes): 22

- Data Types of Features: All 22 features are of type object, which suggests that the variables are likely categorical or text-based, requiring further processing for numerical analysis.

This understanding lays the foundation for the next steps in data analysis and processing

B. Quality Investigation

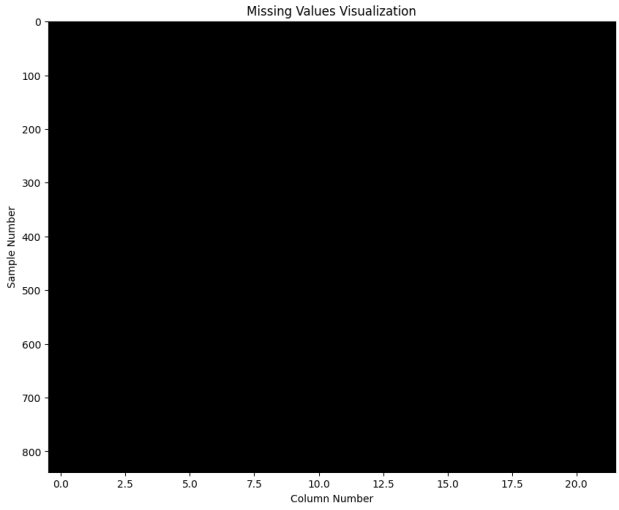
2.1. Duplicates

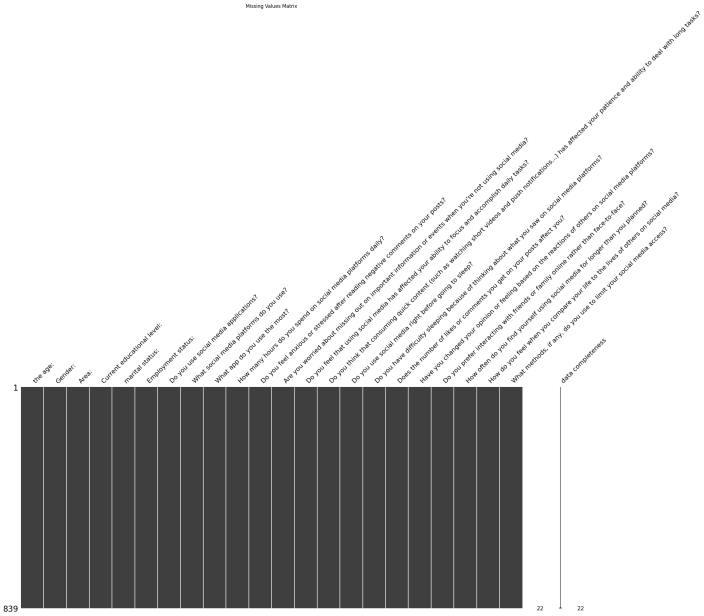
We conducted a **Duplicate Values Analysis** using the duplicated() function from the **Pandas** library, which helps improve data quality by detecting and removing duplicates, contributing to more accurate results when analyzing user behavior and perspectives. However, the result indicated that we do not have any duplicate data.

2.2. Missing values

2.2.1. Per sample

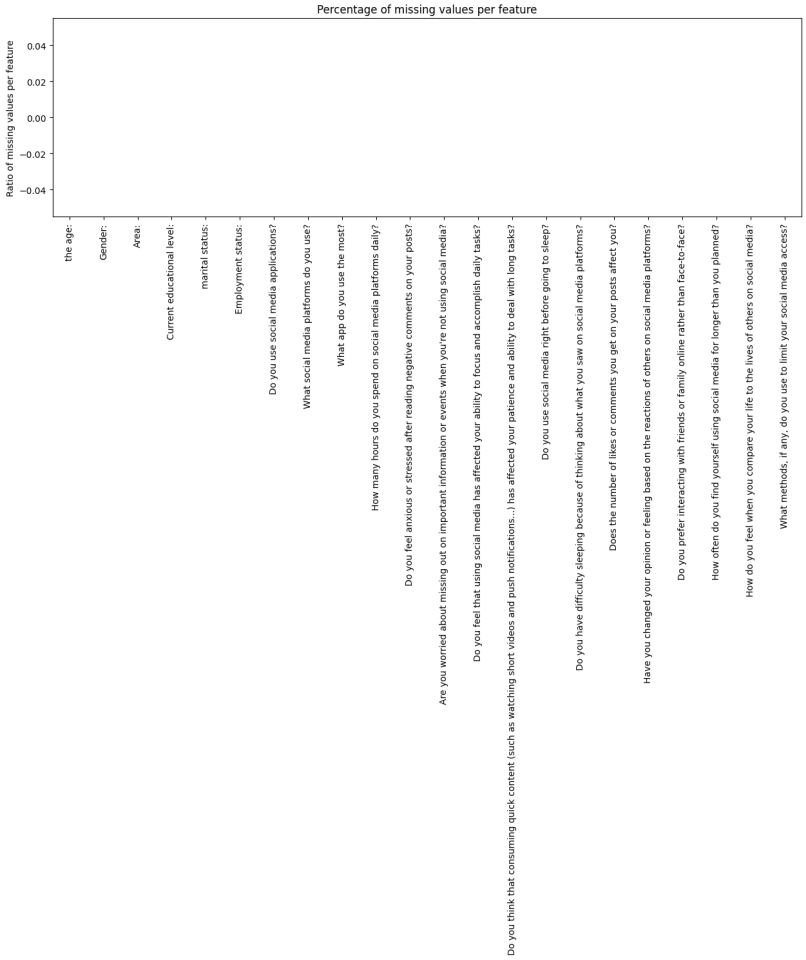
Matplotlib provided us with an overview of the distribution of missing values, showing that there **are no missing values present**. The result was similar with the **Missingno missing values matrix**, which offered detailed information on whether there are missing values in each feature from the generated plot.





2.2.2. Per Feature

Additionally, using Matplotlib for plotting, we analyzed the features to determine if there were any missing values. A bar chart was created to display the percentage of missing values for each feature, but no missing values were found in the features



2.3. Unwanted entries and recording errors

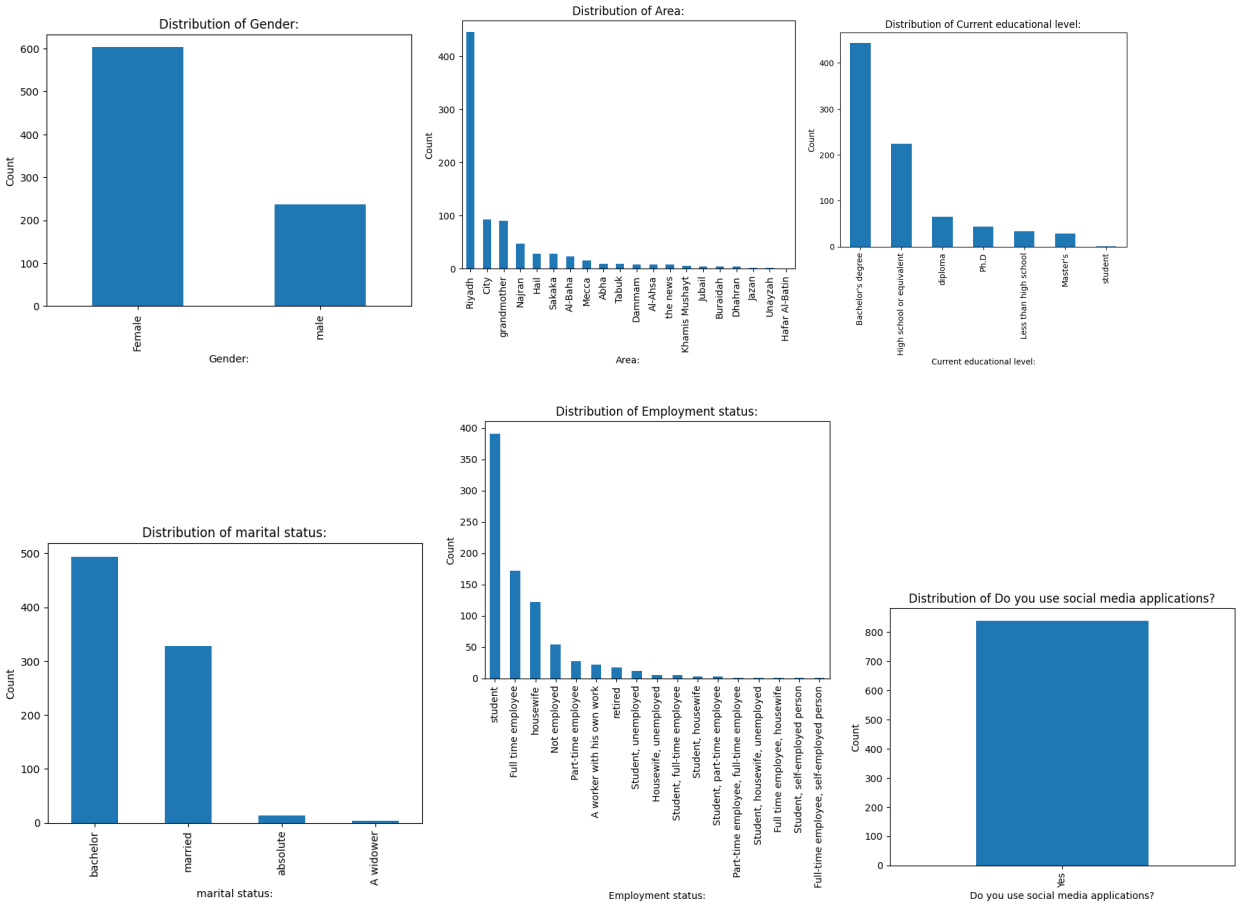
2.3.1. Numerical features

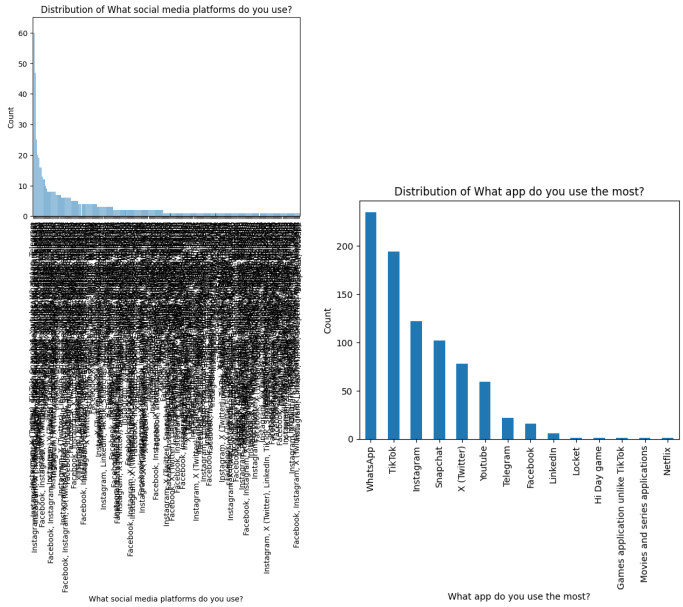
The plot shows the distribution of data for each feature separately, with each column representing a specific feature. Most features exhibit a similar distribution, indicating common behaviors among social media users. However, there are clear differences in the first three columns, which show notable variations that may suggest individual usage habits. Therefore, we can conclude that the similar distribution reflects shared experiences among users, while the differences indicate

varying individual influences.

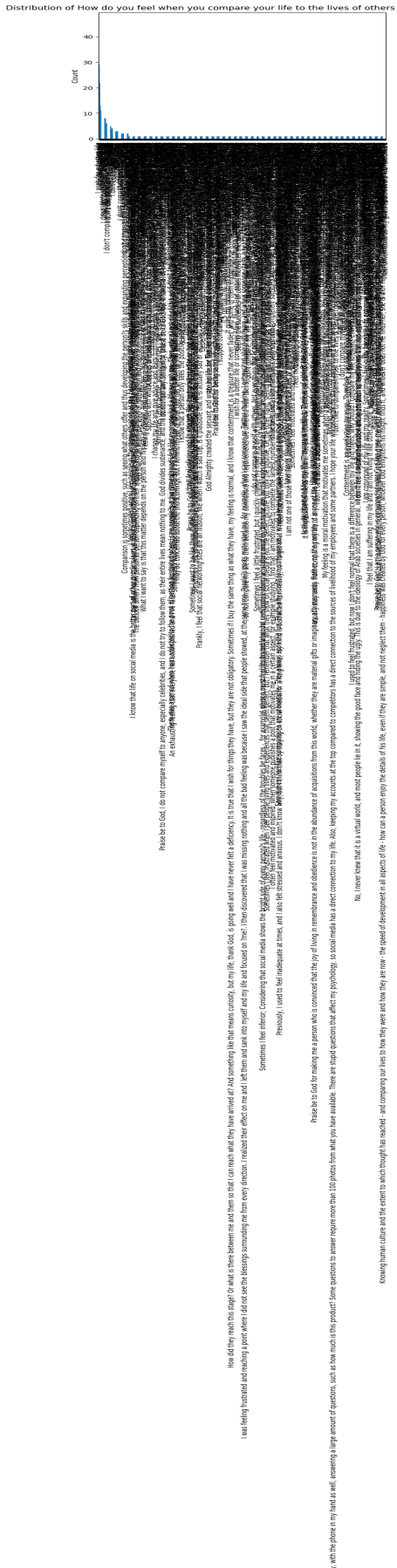
2.3.2. Non-numerical features

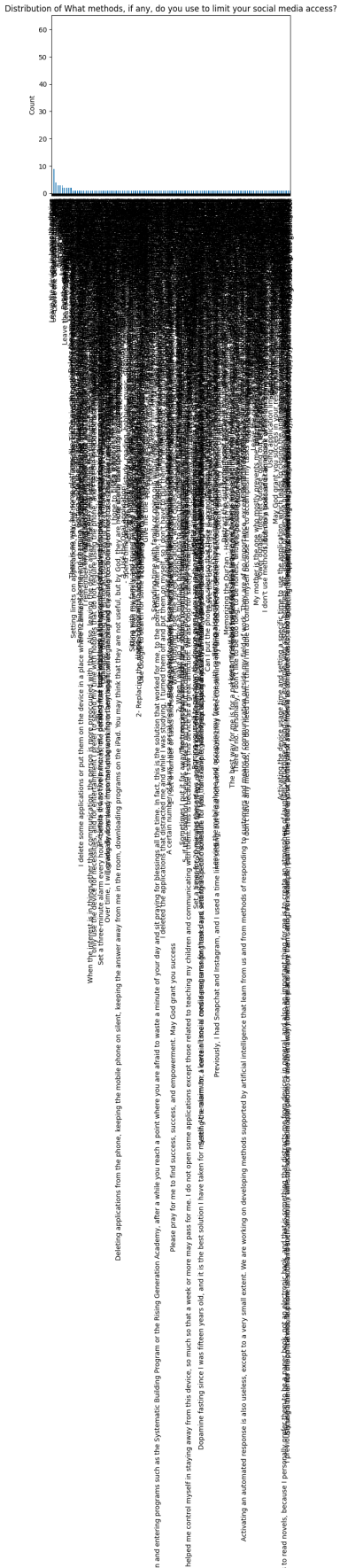
Now, when we look for missing values in non-numerical features, we are checking for gaps in categorical or text-based columns, which might require special handling (like encoding or imputing) to fill or remove.




Distribution of What social media platforms do you use?
Count
What social media platforms do you use?
Distribution of What app do you use the most?
Count
WhatsApp
TikTok
Instagram
Snapchat
X (Twitter)
Youtube
Telegram
Facebook
Linkedin
Locket
Hi Day game
Games application unlike TikTok
Movies and series applications
Netflix
What app do you use the most?

King Saud University
College of Computer and Information Sciences
Information Technology Department





The charts will illustrate the distribution of unique values for each non-numerical feature, allowing you to identify rare categories or unexpected values and assess the quality of the data

2.4. Conclusion of quality investigation

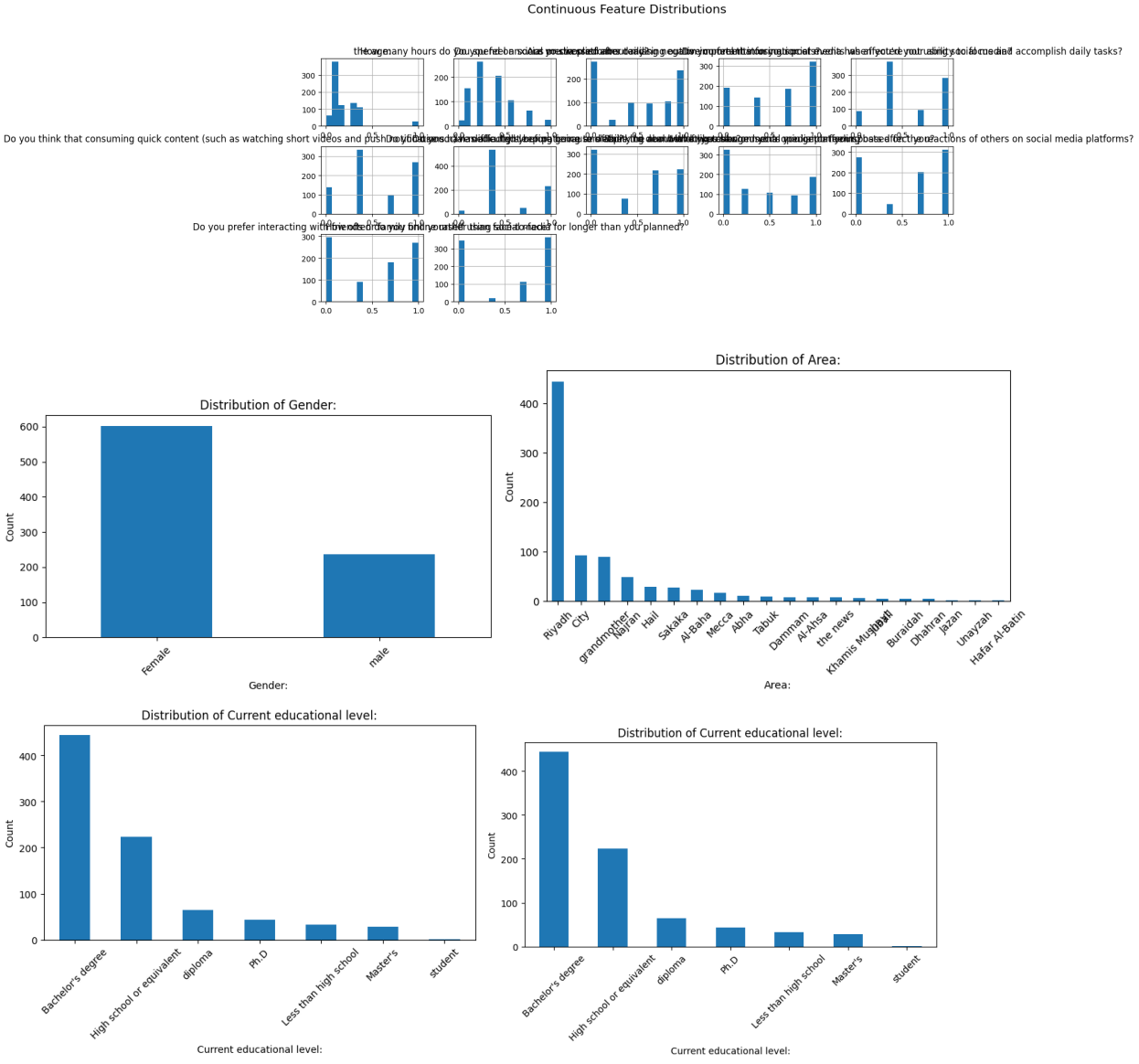
By the end of this second investigation, we should have gained a clearer understanding of the overall quality of our dataset. We examined duplicates, missing values, and unwanted or erroneous entries, ensuring a more accurate and reliable dataset for analysis. However, it's important to note that we have not yet addressed strategies for handling the remaining missing values or outliers, which will be crucial steps in refining the dataset further.

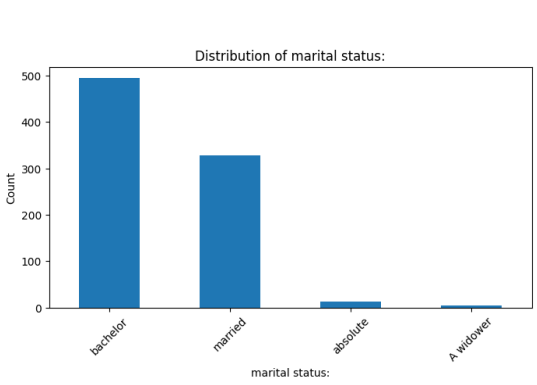
- **C. Content Investigation:**
  In this section, we will conduct an in-depth exploration of the feature values in the dataset to understand how different features interact with each other. This analysis aims to identify significant relationships between features, uncover patterns and trends, and detect outliers. Through this exploration, we will be able to pinpoint which features have a greater impact on the target outcomes, contributing to improved analysis and modeling strategies. This process will also help guide data processing decisions and enhance the overall performance of the models

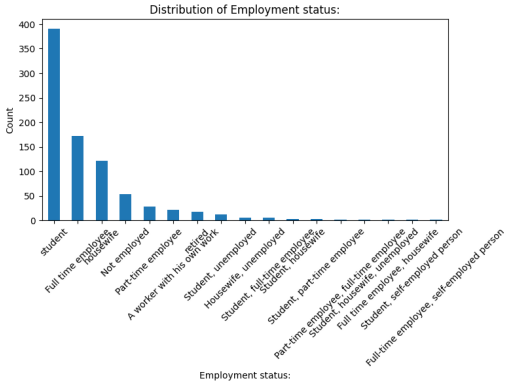
3.1. Feature distribution

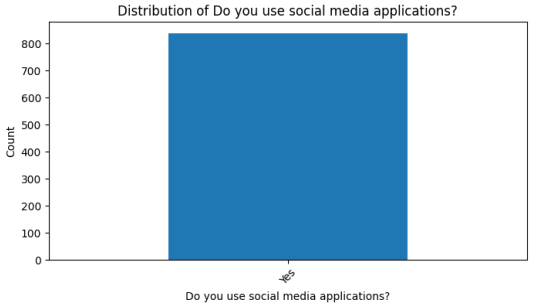
we used histograms and bar charts to visualize the distributions of continuous and discrete features in the data. The results revealed a higher number of females compared to males, with participants primarily concentrated in Riyadh and nearly equal representation from Jeddah and Madinah. Additionally, WhatsApp and TikTok were the most commonly used

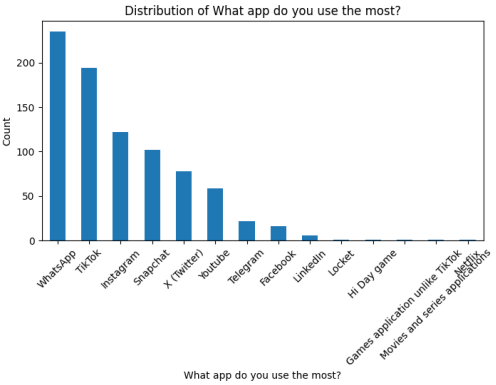
applications among participants, reflecting clear preferences in social media usage. Furthermore, the analysis showed that the majority of participants held a Bachelor's degree, indicating a high level of educational attainment. These findings enhance the understanding of the sample's characteristics and trends, aiding in guiding future analyses.

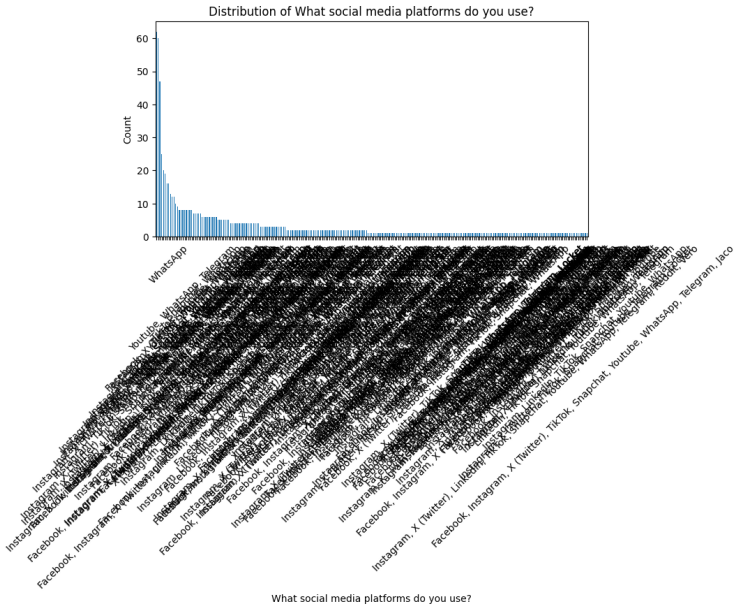



Distribution of marital status:
Count
marital status:
bachelor
married
absolute
A widower

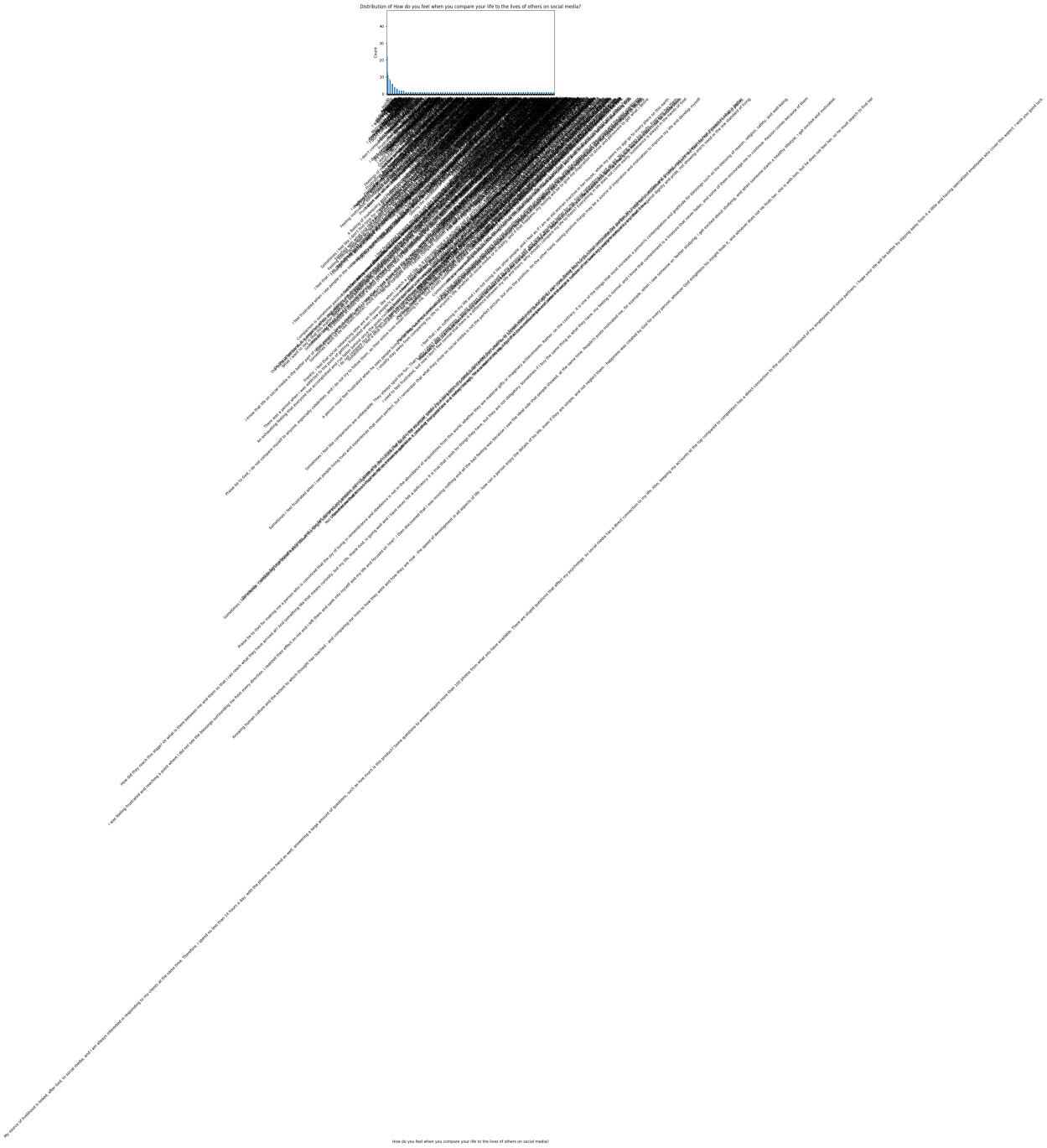

Distribution of Employment status:
Count
Employment status:
student
Full time employee
housewife
Not employed
Part-time employee
A worker with his own work
retired
Student, unemployed
Housewife, unemployed
Student, full-time employee
Student, housewife
Student, part-time employee
Part-time employee, full-time employee
Student, housewife, unemployed
Full time employee, housewife
Student, self-employed person
Full-time employee , self-employed person


Distribution of Do you use social media applications?
Count
Do you use social media applications?
Yes

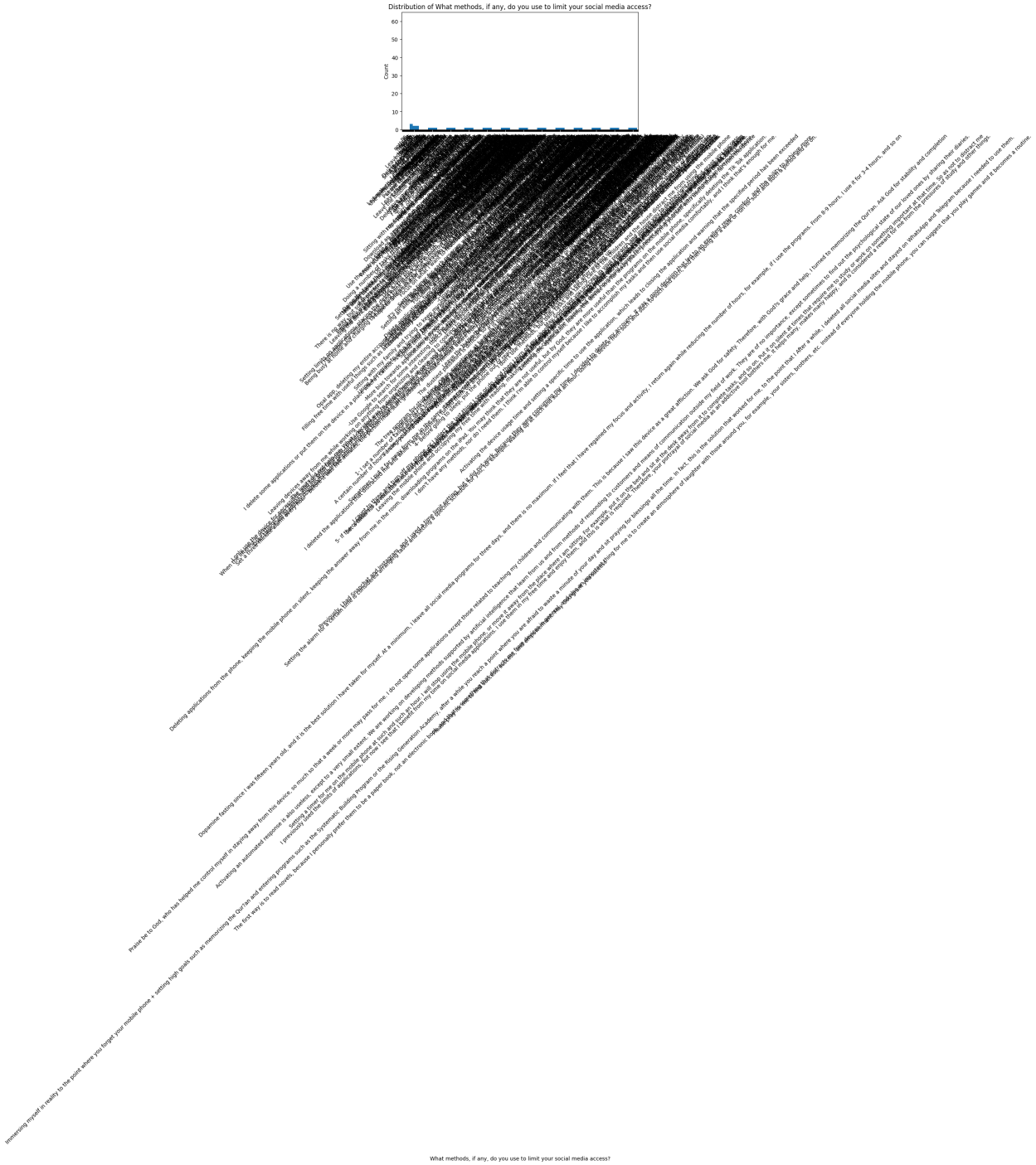

Distribution of What app do you use the most?
Count
What app do you use the most?
WhatsApp
TikTok
Instagram
Snapchat
X (Twitter)
Youtube
Telegram
Facebook
LinkedIn
Locket
Hi Day game
Games application unlike TikTok
Netflix
Movies and series applications


Distribution of What social media platforms do you use?
Count
What social media platforms do you use?
WhatsApp

King Saud University
College of Computer and Information Sciences
Information Technology Department

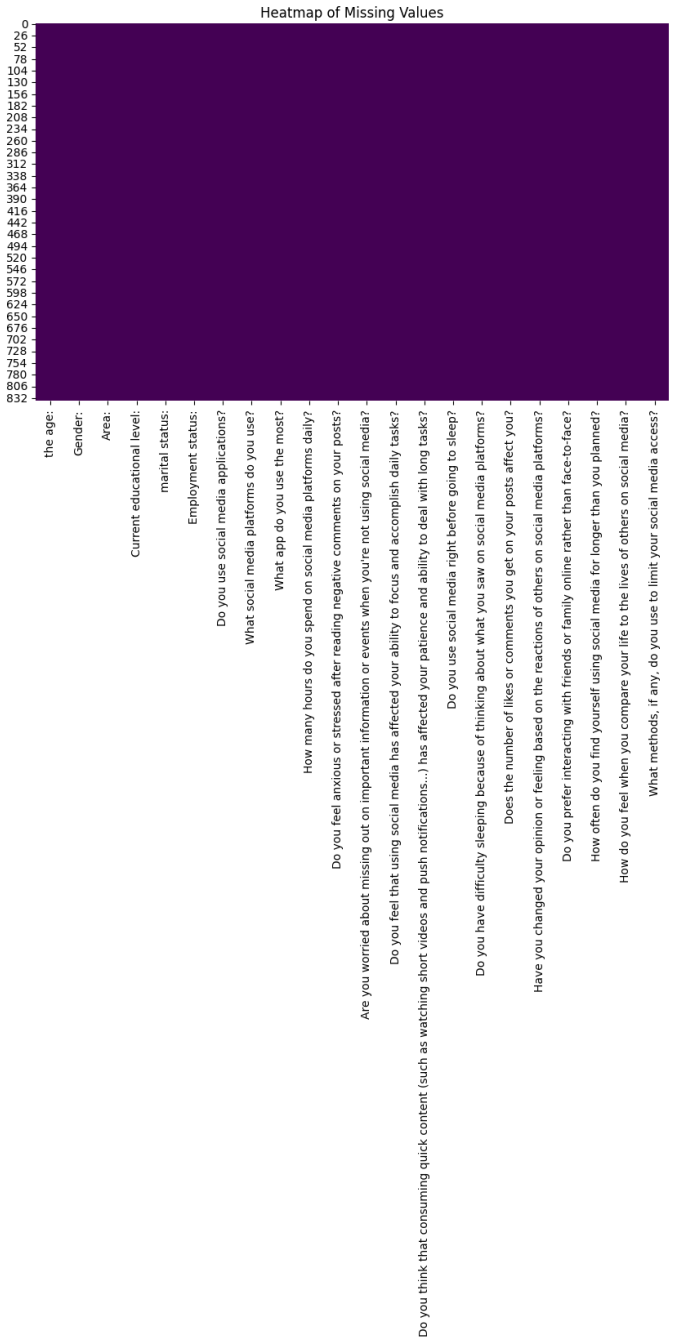


King Saud University
College of Computer and Information Sciences
Information Technology Department



3.2. Feature patterns

The "Feature Patterns" section aims to explore the relationships and patterns within the data, such as correlations and repetitive behaviors, to guide feature engineering and modeling processes. Libraries like **Pandas** were used for data analysis, while **Matplotlib** and **Seaborn** were utilized for visualization. The heatmap revealed that there are no missing values, enhancing the reliability of the data. These insights contribute to a better understanding of the dataset and effectively guide future analyses.



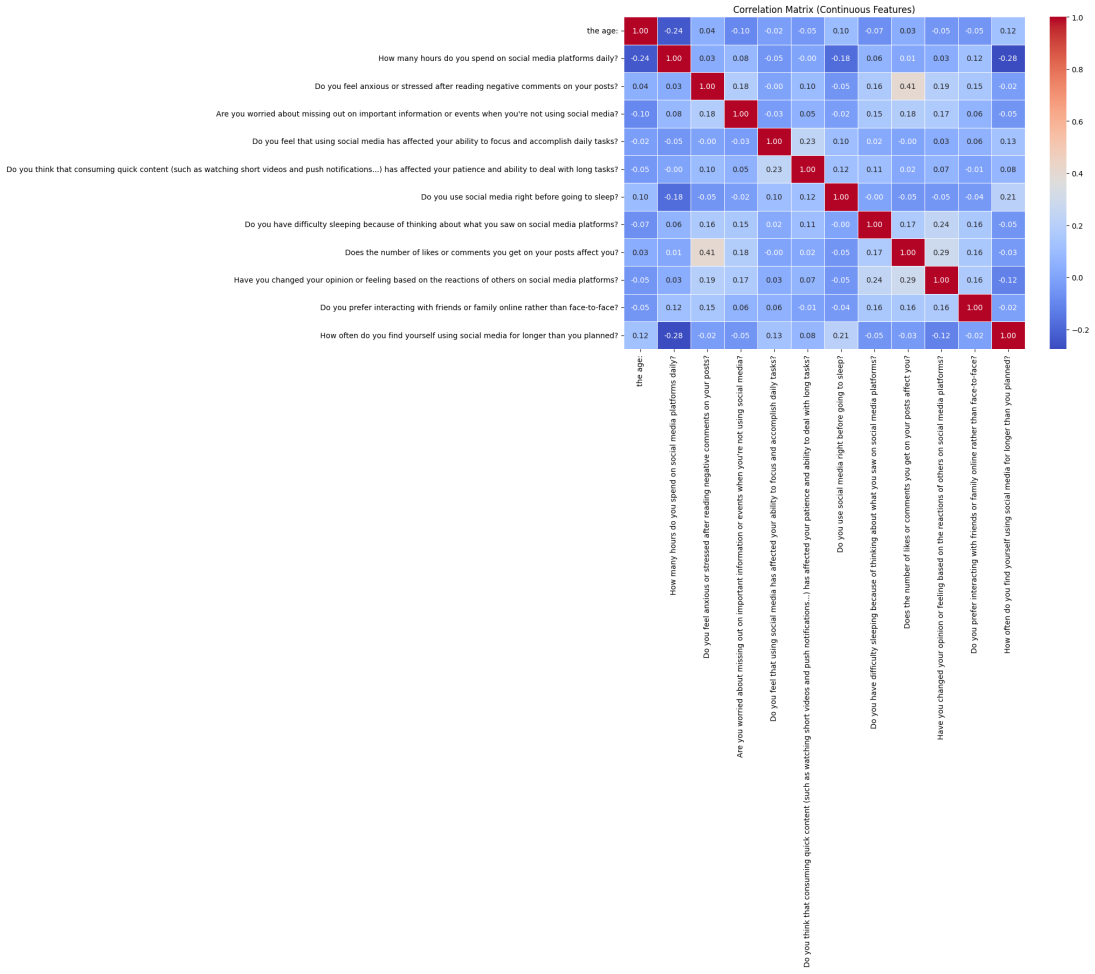
3.3. Feature relationships

The objective of this analysis is to explore the relationships between continuous features in the dataset by calculating their correlations.

Understanding these correlations is essential for gaining insights into how different variables interact with one another.

The Pandas library is utilized for data manipulation and to compute the correlation matrix using the `.corr()` function. Matplotlib is employed for visualizing the correlation matrix, while Seaborn is used to create a heatmap that enhances the visual representation of the correlation matrix.

The correlation matrix helps quickly identify the strength and direction of relationships between numerical features. The heatmap visually reflects these correlations, making it easy to spot strongly correlated pairs. This information is valuable for guiding future analyses and influencing model selection, highlighting the importance of understanding feature interactions for developing effective predictive models.



3.4. Conclusion of content investigation

The analysis of outliers in the dataset reveals valuable insights into participants' responses. By applying the interquartile range (IQR) method, we calculated the number of outliers for each numerical question in the survey.

The results indicate the following counts of outliers for specific questions:

- Hours spent on social media daily: 110 rows with outliers.

- Concern about missing important information while not using social media: 143 rows with outliers.

- Impact of social media on focus and daily tasks: 87 rows with outliers.

- Other questions, including those related to content consumption and interaction preferences, showed no outliers.

In total, there are 340 rows with outliers across the dataset. This information is crucial as it highlights the variability in responses and may indicate participants with extreme opinions or behaviors regarding social media usage. Addressing these outliers will be important for ensuring the robustness of future analyses and interpretations of the data.

- **Statistical Summaries:**

In this analysis, we used the **Pandas** library to compute summary statistics for a dataset containing survey responses. The primary goal was to gain insights into the numerical features of the dataset, which is essential for understanding underlying trends and patterns in the data. Additionally, we utilized the **NumPy** library, which supports multidimensional arrays and provides advanced mathematical functions, enhancing the efficiency of our analysis.

The dataset analyzed using the `summary_stats()` function reveals that the majority of respondents are young adults aged 18-24, with a significant skew towards feminine gender identity. Most participants are from Riyadh, indicating a strong regional concentration, and the sample is highly educated, with many holding a Bachelor's degree. Additionally, a large portion of respondents are single, and many are students. Social media usage is prevalent, with a majority of respondents actively using it, particularly favoring WhatsApp as the most frequently used platform. Overall, the dataset is characterized by young, educated, feminine individuals primarily residing in Riyadh.

- **The Variance**

Objective of Variance:

The goal of variance analysis is to measure the spread or dispersion of a particular set of data points, which helps in understanding how values differ from the mean. This information is essential for assessing the diversity of the data and guiding decisions on appropriate analyses and modeling techniques.

Results:

```
Variance for all numerical columns:
the age:                                                                                                                                0.037617
How many hours do you spend on social media platforms daily?                                                                            0.052311
Do you feel anxious or stressed after reading negative comments on your posts?                                                          0.168641
Are you worried about missing out on important information or events when you're not using social media?                                0.155843
Do you feel that using social media has affected your ability to focus and accomplish daily tasks?                                      0.121913
Do you think that consuming quick content (such as watching short videos and push notifications...) has affected your patience and ability to deal with long tasks?    0.135104
Do you use social media right before going to sleep?                                                                                    0.094815
Do you have difficulty sleeping because of thinking about what you saw on social media platforms?                                       0.171071
Does the number of likes or comments you get on your posts affect you?                                                                  0.160966
Have you changed your opinion or feeling based on the reactions of others on social media platforms?                                    0.180879
Do you prefer interacting with friends or family online rather than face-to-face?                                                       0.178158
How often do you find yourself using social media for longer than you planned?                                                          0.215165
dtype: float64
```

Variance of Age: Moderate, indicating diversity among age groups, with a significant representation in the 18-24 age range.

Variance of Gender: Very low, due to a skewed distribution towards females (606 out of 851 respondents).

Variance of City: Moderate, as there are 20 unique areas, but Riyadh dominates the distribution.

Variance of Social Media Usage: Low, since most respondents use social media, particularly WhatsApp.

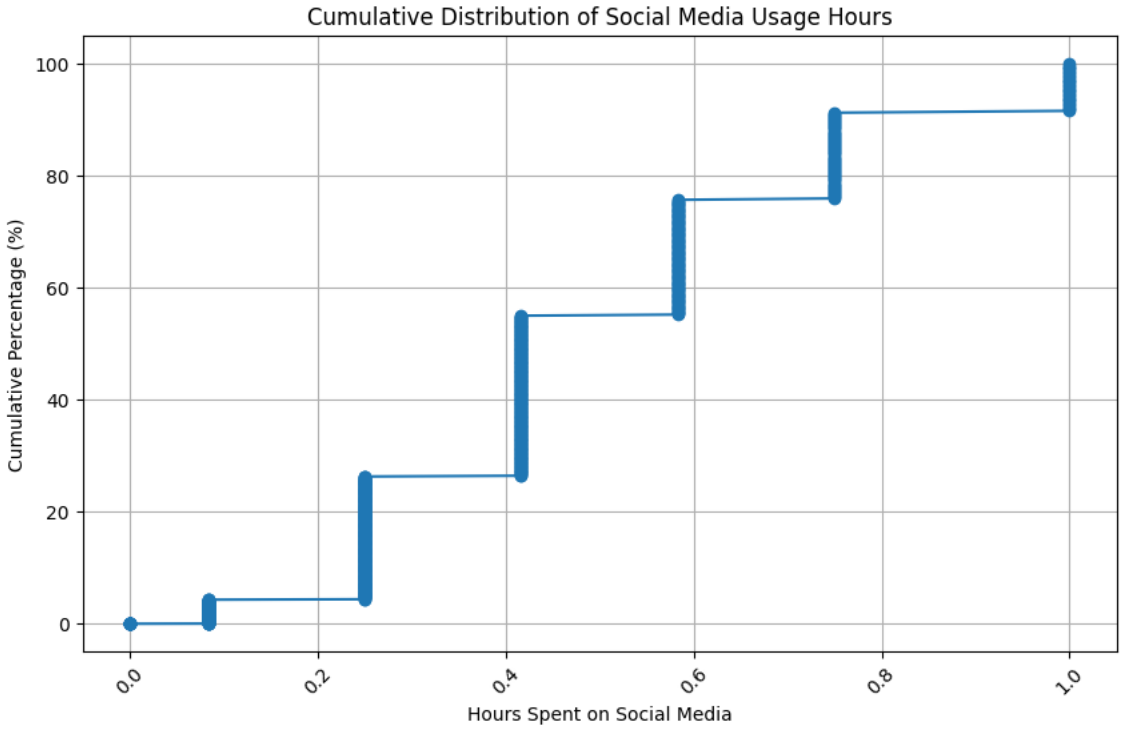
Variance in Other Questions: Shows moderate variance in responses regarding anxiety and focus, with the highest variance in spending longer on social media than planned.

Tools Used: To analyze the variance, the **Pandas** library in Python was utilized, where the dataset was loaded, and the variance for all numerical columns was calculated using the var() function.

- **Data Visualization**

In these analyses, the **Pandas** library was used to load and process survey data on the impact of social media on mental health. Two main analyses were conducted using **Matplotlib** and **Seaborn** libraries for data visualization.

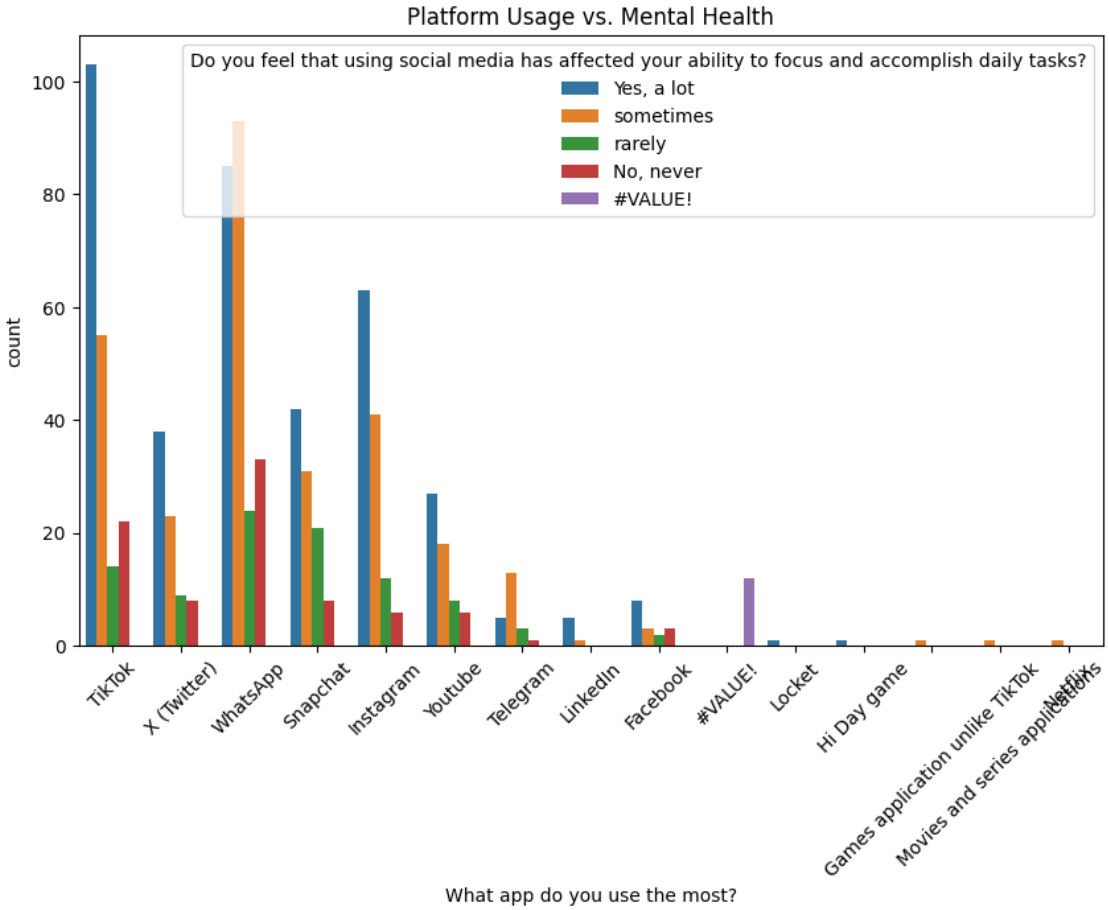
The first analysis focused on the relationship between the number of hours individuals spend on social media and how it affects their ability to focus. The results showed that the data points were evenly distributed, indicating no clear correlation between the hours spent and the ability to concentrate.



In the second analysis, a bar plot was created to compare the usage of social media platforms and their relationship with mental health outcomes, particularly regarding the ability to focus. The **Pandas** library was used for data processing, while **Matplotlib** and **Seaborn** were employed for visualization.

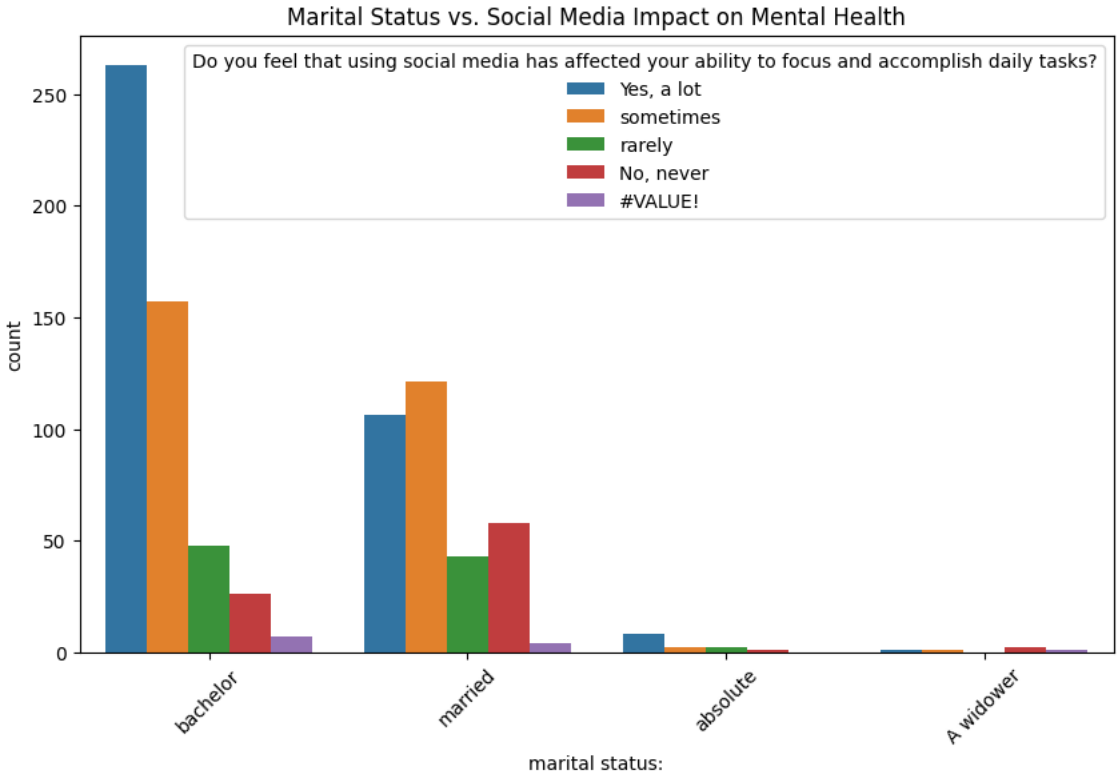
The bar plot displays the most frequently used apps alongside respondents' perceptions of how social media impacts their focus and ability to accomplish daily tasks. The x-axis represents different social media platforms, and the bars are color-coded based on whether users feel that social media affects their concentration.

The analysis reveals that **TikTok** users feel that short videos significantly impact their focus, followed by **WhatsApp** and **Instagram** users. This indicates that certain social media platforms may have a greater influence on individuals' daily tasks than others.

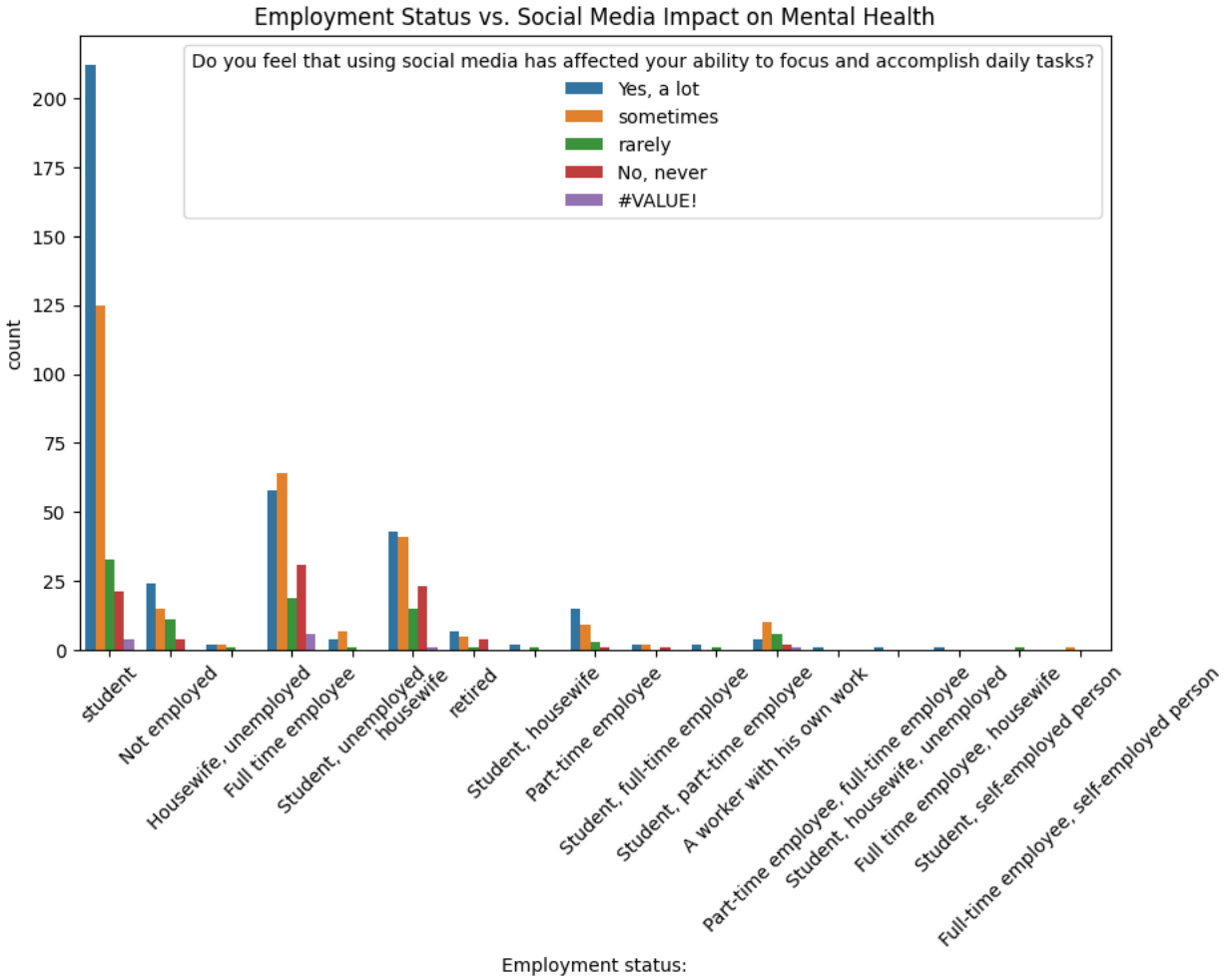


In this third analysis the grouped bar plot illustrates the relationship between respondents' marital status and their perception of how social media affects their focus. The x-axis represents different marital statuses, with the bars colored according to users' perceptions of social media's impact on their concentration. This visualization aims to highlight significant differences in perceptions across various relationship statuses, aiding in the understanding of how the effects of social media

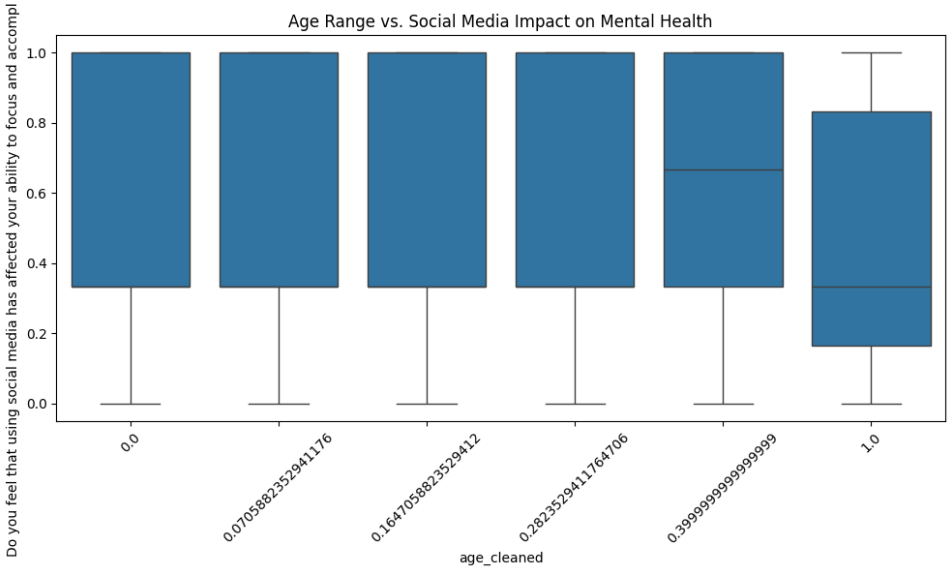
on mental health may vary based on marital status, whether respondents are single, married, or in a relationship.



The grouped bar plot analyzes the impact of social media on mental health across different employment statuses—employed, unemployed, and students. It shows that students are the most affected group, highlighting a significant correlation between social media use and mental health challenges in this demographic. This visualization offers key insights into how social media influences concentration and task completion based on employment status.



The box plot illustrates the relationship between age groups and the impact of social media on mental health. The results indicate that the first four age categories show similar levels of impact, while the later age groups differ from each other. This visualization highlights the significance of age differences in understanding how social media affects the ability to focus and accomplish daily tasks.


Age Range vs. Social Media Impact on Mental Health
Do you feel that using social media has affected your ability to focus and accompl
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.07058823529411764
0.16470588235294112
0.2823529411764706
0.3999999999999999
1.0
age_cleaned

Logbook Entry: Data Processing and Cleaning

- **Data Transformation**

  Step 1: Replacing Inconsistent Gender Values

  - In this step, the value "feminine" in the "Gender" column was replaced with "Female." This change was necessary to ensure consistent and standardized values, which are crucial for accurate analysis and reporting. Using "Female" instead of "feminine" guarantees uniformity across the dataset, facilitating better interpretation of gender-related data. The implementation involved utilizing a code snippet that applied the `replace` function to update the inconsistent value in the dataset**.**

```
# Replace "feminine" with "Female" in the Gender column
survey_data['Gender:'] = survey_data['Gender:'].replace("feminine", "Female")
```

Step 2: Correcting Incorrect City Names

In this step, several incorrect city names in the "Area" column were corrected: "grandmother" was changed to "Jeddah," "the news" was changed to "Khobar," and "City" was changed to "Madinah."

This correction was necessary because accurate location data is essential for geographic analysis and understanding regional differences in survey responses. By ensuring that the city names are accurate, we prevent potential misinterpretation of the data. The implementation involved using a code snippet that utilized the replace function to update the incorrect values in the dataset

```
# Replace incorrect city names in the City column
survey_data['Area:'] = survey_data['Area:'].replace({
    "grandmother": "Jeddah",
    "the news": "Khobar",
    "City": "Madinah"
})
```

- **Response Conversion to Numerical Range (1-5)**

  In this step, we standardized categorical survey responses by converting them into a numerical range from 1 to 5, facilitating quantitative analysis. We first identified the relevant columns for conversion and defined a mapping dictionary to associate each response with a specific numerical value. For instance, responses like "Yes, always" were mapped to 1, while "No, never" was mapped to 5. We then applied this mapping to the specified columns in the dataset to replace categorical responses with their corresponding numerical values. Finally, we displayed a sample of the updated dataset to verify the conversion's accuracy. This process is essential for enabling statistical analysis and machine learning applications, as they require numerical input for effective data processing.

- **Cleaning and Converting Hours Data**

  In this step, we cleaned and standardized the "Hours" data in our dataset, which contained various inconsistent entries. A function named clean_hours was defined to handle different formats. It extracted numeric values from date-like entries (e.g., "4-Mar"), replaced invalid entries like "#VALUE!" with None, converted "Less than an hour" to 0.5 hours, assigned 13 to "12 hours or more," and extracted hours from general strings like "2 hours." The function was then applied to create a new column, "Hours_Cleaned." Rows with missing values in this new column were dropped to ensure a clean dataset. Finally, we displayed the unique cleaned hour values to verify the process's effectiveness. This cleaning was crucial for preparing the dataset for accurate analysis and statistical modeling.