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 (Primary Data)

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

Logbook Entry: Exploratory Data Analysis (EDA)

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

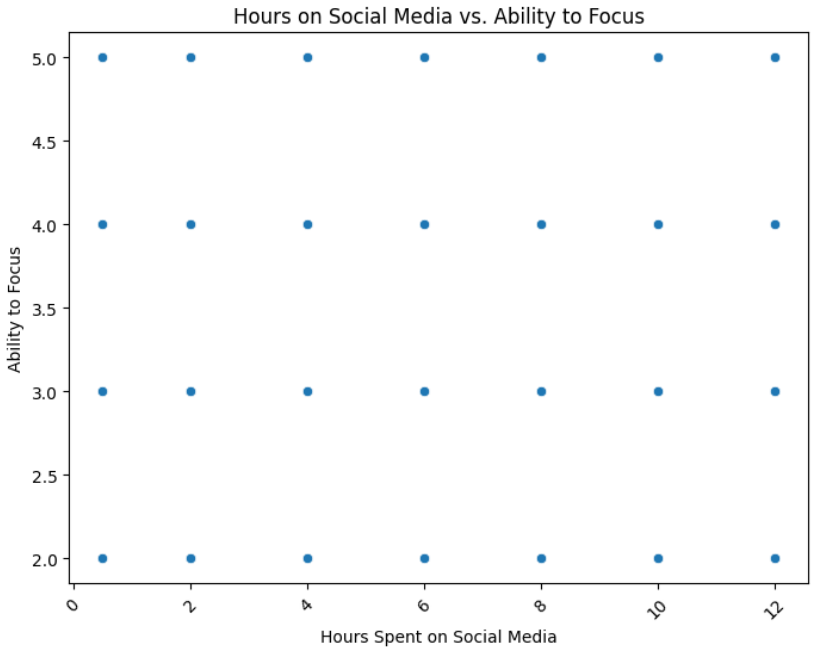
In these analyses, the **Pandas** library was used to load and process survey data on the impact of social media on mental health. Two main analyses were conducted using **Matplotlib** and **Seaborn** libraries for data visualization.

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.

The second analysis explored the relationship between the frequency of social comparisons on social media and mental health outcomes. A bar plot was used to analyze individuals' feelings when making social comparisons and its impact on their ability to focus. The findings suggested a particular pattern in how social comparisons influence mental health, highlighting the need for further study to understand these dynamics better.

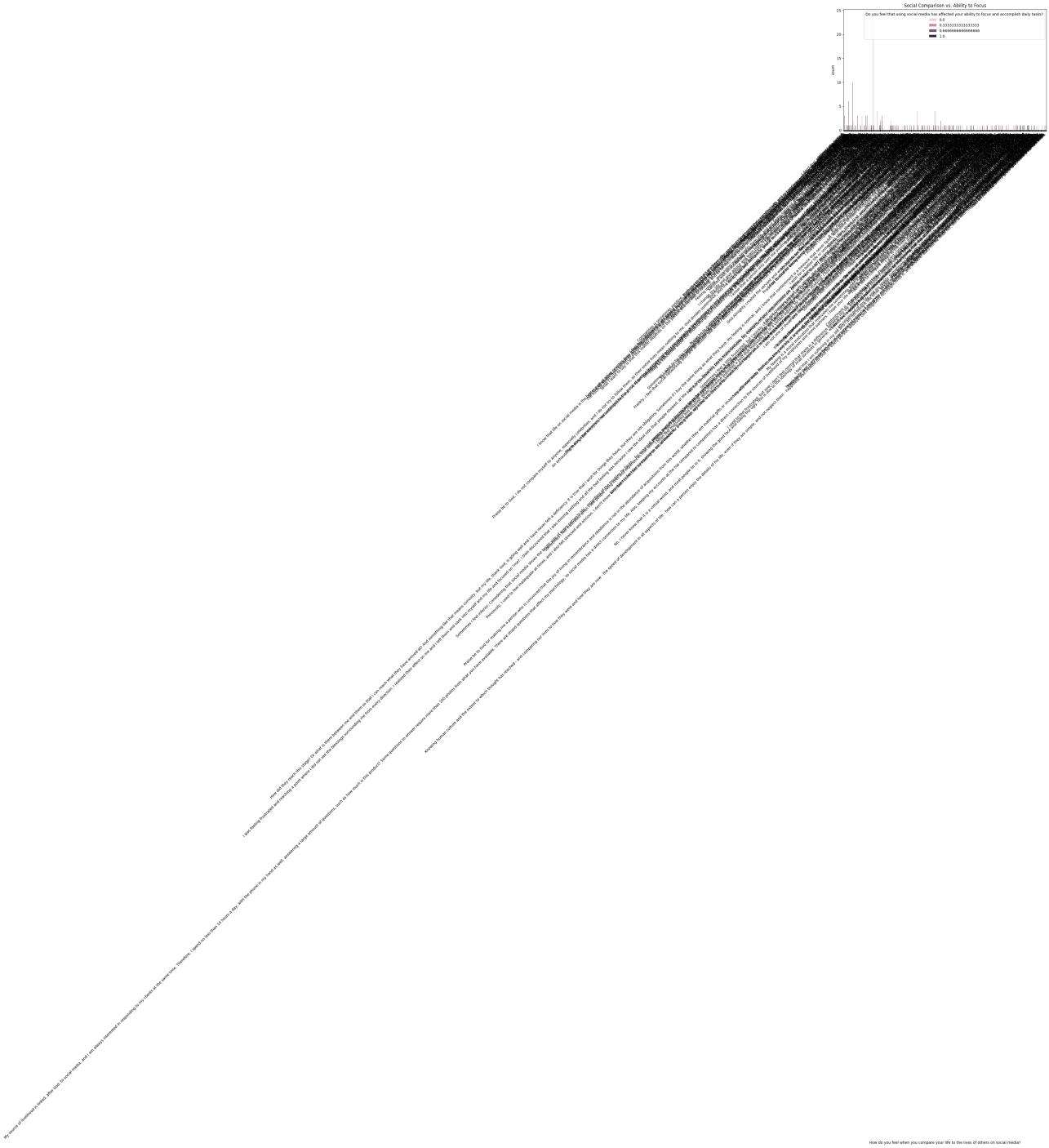
These analyses contribute to understanding the potential effects of social media usage on mental health and individual behavior.

first analysis:



The second analysis:

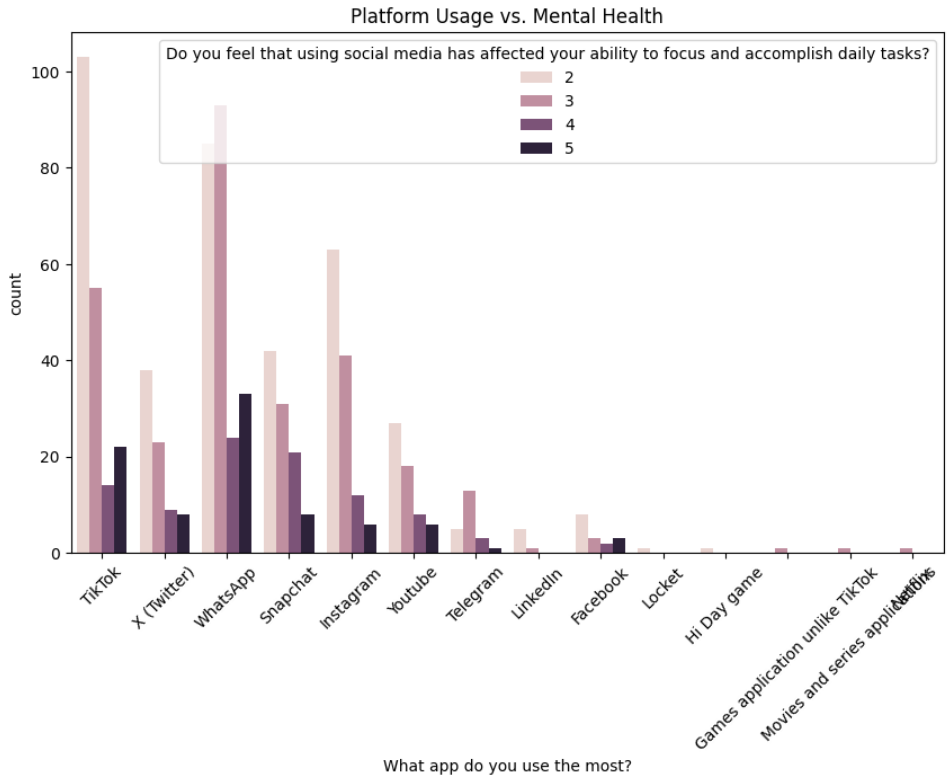
King Saud University
College of Computer and Information Sciences
Information Technology Department



In thisTheared 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.

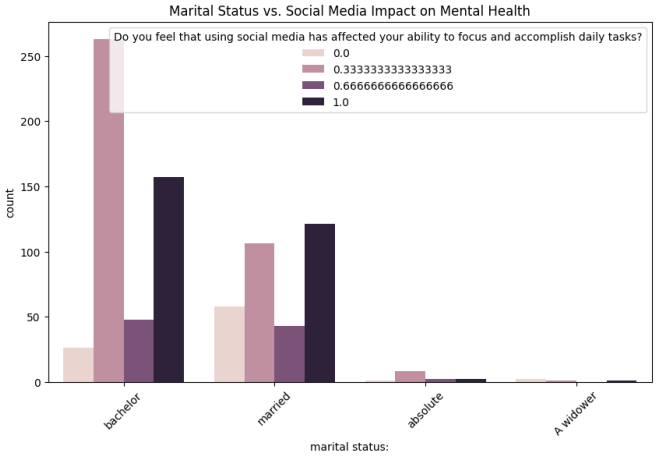
The bar plot displays the most frequently used apps alongside respondents' perceptions of how social media impacts their focus and ability to accomplish daily tasks. The x-axis represents different social media platforms, and the bars are color-coded based on whether users feel that social media affects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

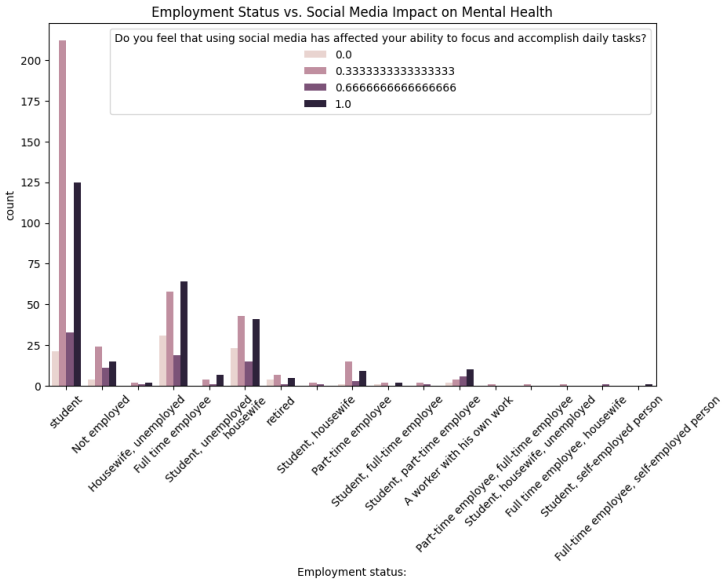


In this Fourth analysis the grouped bar plot illustrates the relationship between respondents' marital status and their perception of how social media affects their focus. The x-axis represents different marital statuses, with the bars colored according to users' perceptions of social media's impact 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.

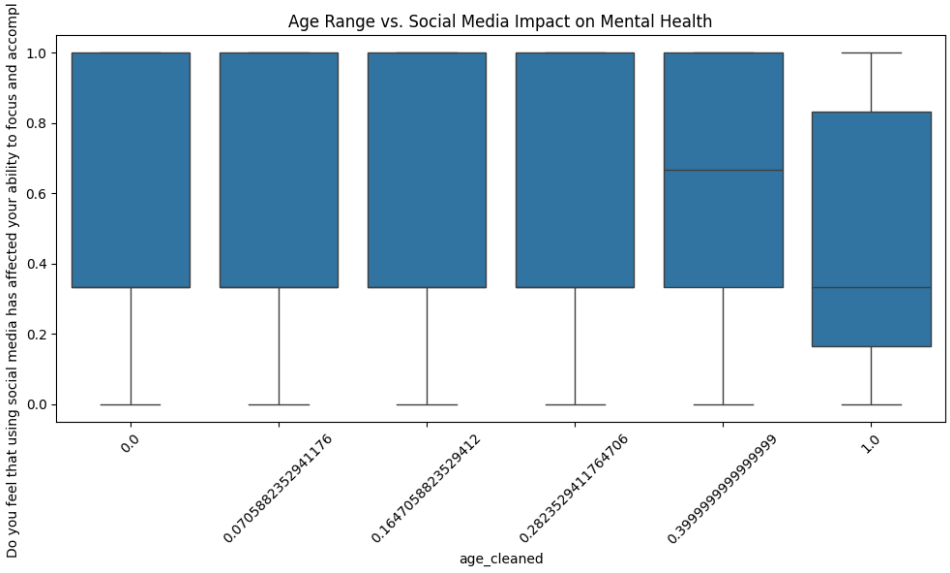
s



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.



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

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.