Data Analysis Project: Social Media and Mental Health

Serena Hinton 2024

Introduction

The Google Data Analytics Professional Certificate Capstone project is a case study. The program offered two tracks that offered either a case study provided by the program or the option to create our own. I chose to create my own since this was my first project and I wanted the practice.

All of my code can be found on Github.

In this project I used Python. This also serves as my first complete Python project.

The project covers each step of the data analysis process: Ask, Prepare, Process, Analyze, Share, and Act.

Scenario

You are a junior data analyst and have been asked to lead a project for a brand new client—this will involve everything from defining the business task all the way through presenting your data-driven recommendations. You will choose the topic, ask the right questions, identify a dataset and ensure its integrity, conduct analysis, create compelling data visualizations, and prepare a presentation.

About the Company

SocialTech Enterprises, Inc. is a prominent technology company specializing in social media platforms and online communities. Founded a decade ago by entrepreneur Sarah Mitchell, the company has been at the forefront of providing innovative digital solutions to connect people worldwide. With a user base exceeding millions, the company's platforms have played a significant role in shaping online interactions and the way people communicate in the digital age.

Recently, SocialTech Enterprises has been facing increasing scrutiny and public concern over the potential impact of their platforms on users' mental health and well-being. There have been media reports, academic studies, and user feedback suggesting that prolonged engagement with their platforms may contribute to issues such as anxiety, low self-esteem, and even symptoms of ADHD and depression.

Amidst these concerns, Sarah Mitchell, the CEO of SocialTech Enterprises, has taken a proactive approach to address this challenge. She believes that technology, when used thoughtfully, can enhance lives and well-being.

The goal is to provide SocialTech Enterprises with data-informed insights and actionable recommendations that will enable the company to *create and maintain an online environment that promotes better mental health and well-being among its users.*

---

1. Ask

Business Task

What is the relationship between social media use and user psychological well-being?

Key Stakeholders

- SocialTech Enterprises: prominent technology company specializing in social media platforms and online communities with a user base exceeding millions.
- Sarah Mitchell—CEO of SocialTech Enterprises, Inc.

2. Prepare

The public data set here will be used. The data has been made available by SouvikAhmed071 and Muhesena from Kaggle with this license.

I used Python to process and visualize the data.

Columns include:

- Demographics: Age, Occupation, Organization Affiliation, Gender, Relationship Status
- Mental Health Screening Questions for ADHD, Anxiety, Depression, and Self-esteem
- Number of Hours Spent on Social Media
- Types of Social Media Platforms Used

Mental Health Screening Questions scored with Likert Scale: 1(low)–5(high) with 60 points being the highest possible total score:

1. ADHD: 4 Questions, 20 point possible
    1.1. How often do you find yourself using Social media without a specific purpose?
    1.2. How often do you get distracted by Social media when you are busy doing something?
    1.3. On a scale of 1 to 5, how easily distracted are you?
    1.4. Do you find it difficult to concentrate on things?
2. Anxiety: 2 Questions, 10 point possible

    2.1. Do you feel restless if you haven't used Social media in a while?
    2.2. On a scale of 1 to 5, how much are you bothered by worries?'
3. Depression: 3 Questions, 15 points possible
    3.1. How often do you feel depressed or down?':' Depression Q1',
    3.2. On a scale of 1 to 5, how frequently does your interest in daily activities fluctuate?': 'Depression Q2',
    3.3. On a scale of 1 to 5, how often do you face issues regarding sleep?':' Depression Q3'
4. Self-Esteem: 3 Questions, 15 points possible
    4.1. On a scale of 1-5, how often do you compare yourself to other successful people through the use of social media?':' Self Esteem Q1',
    4.2. Following the previous question, how do you feel about these comparisons, generally speaking?':' Self Esteem Q2',
    4.3. How often do you look to seek validation from features of social media?': 'Self Esteem Q3',

*The dataset lacks predefined intervals for categorizing scores into mild, moderate, or severe categories based on the total score. To provide clarity, I will designate scores below 20 as mild, scores from 21 to 40 as moderate, and scores from 41 to 60 as severe.

---

2. Process

2.1 Rename the Columns

Shortening the column names will make it easier to read.

```
# Printing first 5 entries in the data set
data.head()
```

:

	Timestamp	1. What is your age?	2. Gender	3. Relationship Status	4. Occupation Status	5. What type of organizations are you affiliated with?	6. Do you use social media?	7. What social media platforms do you commonly use?	8. W is th aver time sper on soci med ever day?
0	4/18/2022 19:18:47	21.0	Male	In a relationship	University Student	University	Yes	Facebook, Twitter, Instagram, YouTube,	Betv 2 an hour

```
#rename columns for necessary information
data.rename(columns = {'1. What is your age?':'Age','2. Gender':'Gender','3. Relationship Status':'Relationship Status',
                       '4. Occupation Status':'Occupation',
                       '5. What type of organizations are you affiliated with?':'Affiliations',
                       '6. Do you use social media?':'Social Media User?',
                       '7. What social media platforms do you commonly use?':'Platforms Used',
                       '8. What is the average time you spend on social media every day?':'Hours Per Day',
                       '9. How often do you find yourself using Social media without a specific purpose?':'ADHD Q1',
                       '10. How often do you get distracted by Social media when you are busy doing something?':'ADHD Q2',
                       "11. Do you feel restless if you haven't used Social media in a while?":'Anxiety Q1',
```

2.2 Rearrange Columns

Make ADHD and Anxiety questions sequential for clarity.

```
# rearranging ADHD and anxiety question columns so that they are sequential

titles[11], titles[12] = titles[12], titles[11]
titles[12], titles[14] = titles[14], titles[12]
titles[13], titles[14] = titles[14], titles[13]
data = data[titles]
titles
```

2.3 Dropping Columns

I will not need timestamp or organization affiliation as it is not relevant to my project

```
#drop columns that are not relevant to this project
to_drop = ['Timestamp',
           'Affiliations']

data.drop(to_drop, inplace=True, axis=1)
```

2.4 Missing Values

```
data = pd.read_csv('/kaggle/input/social-media-and-mental-health/smmh.csv')

# Find missing values
missing_values = data.isnull()

# Display missing values
print("Missing values:")
print(missing_values)
```

2.5 Rename Data Values

2.5.1 Gender

Gender

The "Gender" column has unique "Non-Binary" entries becuase of how the data was collected. Particpants were able to enter string values for Gender. Thus, renaming these columns as "Non-Binary" brings cohesiveness.

```
#List all the unique Gender entries.

Genders = set(data['Gender'])
print(Genders)
```

```
{'Non binary ', 'Female', 'There are others???', 'Male', 'NB', 'Trans',
'unsure ', 'Non-binary', 'Nonbinary '}
```

```
#Combining the unique entries that all fall under the "Non-Binary" category
data.replace('Non-binary','Non-Binary', inplace=True)
data.replace('Nonbinary ','Non-Binary', inplace=True)
data.replace('NB','Non-Binary', inplace=True)
data.replace('Non binary ','Non-Binary', inplace=True)
```

```
Genders = set(data['Gender'])
print(Genders)
```

```
{'Non-Binary', 'Female', 'There are others???', 'Male', 'Trans', 'unsure
'}
```

2.5.2 Hours Per Day

```
# Count unique values in the 'Hours Per Day' column
hour_counts = data['Hours Per Day'].value_counts()

# Print the counts
print("Counts of Hours Spent Per Day:")
print(hour_counts)
```

```
#Setting new value labels
data.loc[data['Hours Per Day'] == 'More than 5 hours', 'Hours Per Day'] = 'M
ore than 5 Hours'
data.loc[data['Hours Per Day'] == 'Between 2 and 3 hours', 'Hours Per Day']
= '2-3 Hours'
data.loc[data['Hours Per Day'] == 'Between 3 and 4 hours', 'Hours Per Day']
= '3-4 Hours'
data.loc[data['Hours Per Day'] == 'Between 1 and 2 hours', 'Hours Per Day']
= '1-2 Hours'
data.loc[data['Hours Per Day'] == 'Between 4 and 5 hours', 'Hours Per Day']
= '4-5 Hours'
data.loc[data['Hours Per Day'] == 'Less than an Hour', 'Hours Per Day'] = 'L
ess than 1 Hour'
```

2.6 Scalar Adjustment

Self-esteem scores operate differently from the other questions. In the other questions, higher scores equal worse mental health, but self-esteem higher scores mean better mental health. I swap the scale to make it consistent.

```
#Setting scale scores
data.loc[data['Self Esteem Q2'] == 1, 'Self Esteem Q2'] = 5
data.loc[data['Self Esteem Q2'] == 2, 'Self Esteem Q2'] = 4
data.loc[data['Self Esteem Q2'] == 3, 'Self Esteem Q2'] = 3
data.loc[data['Self Esteem Q2'] == 4, 'Self Esteem Q2'] = 2
data.loc[data['Self Esteem Q2'] == 5, 'Self Esteem Q2'] = 1
```

2.7 Data Aggregation

2.7.1 Mental Health Screening Questions

The different mental health screening questions are added together for a final score.

```
#Sum scores and add new column

ADHD = ['ADHD Q1', 'ADHD Q2', 'ADHD Q3', 'ADHD Q4']
data['ADHD Score'] = data[ADHD].sum(axis=1)

Anxiety = ['Anxiety Q1', 'Anxiety Q2']
data['Anxiety Score'] = data[Anxiety].sum(axis=1)

SelfEsteem = ['Self Esteem Q1', 'Self Esteem Q2', 'Self Esteem Q3']
data['Self Esteem Score'] = data[SelfEsteem].sum(axis=1)

Depression = ['Depression Q1', 'Depression Q2', 'Depression Q3']
data['Depression Score'] = data[Depression].sum(axis=1)

Total = ['ADHD Score', 'Anxiety Score', 'Self Esteem Score', 'Depression Sco
re']
data['Total Score'] = data[Total].sum(axis=1)
```

2.7.2 Aggregate Social Media Platforms into Categories

```
# Define a function to categorize platforms
def categorize_platforms(platforms_used):
    categories = []
    if any(platform in platforms_used for platform in ['Facebook', 'Twitte
r', 'Discord', 'TikTok']):
        categories.append('Social Networking')
    if any(platform in platforms_used for platform in ['Pinterest', 'YouTub
e', 'Snapchat', 'Instagram']):
        categories.append('Media Sharing')
    if 'Reddit' in platforms_used:
        categories.append('Discussion Forum')
    return ', '.join(categories) if categories else 'Other/Unknown'

# Apply the categorization function to create a new 'Category' column
data['Platforms'] = data['Platforms Used'].apply(categorize_platforms)
```

2.8 Fix User Errors

2.8.1 Social Media User?

There are 3 people who selected ‘no’ for the social media user question but also chose social media platforms that they use—determined to be user error and changing all ‘no’ to ‘yes’. Otherwise, the analysis would be inaccurate.

```
data.loc[data['Social Media User?'] == 'No', 'Social Media User?'] = 'Yes'
```

```
#check that change was made
social_media_user_counts = data['Social Media User?'].value_counts()

# Print the counts
print("Counts of Social Media User?:")
print(social_media_user_counts)
```

2.8.2 Occupation Selection

A portion of participants over age 18 selected ‘school student’ as their occupation. This occupation designation is used for those under 18 as well. Since there is a ‘university student’

occupation available, I switch this group from ‘school student’ to ‘university student’ to get more accurate results for both occupations.

```
# Filter rows where the current occupation is 'School Student' and the age is over 18, and update the 'Occupation' column
data.loc[(data['Occupation'] == 'School Student') & (data['Age'] > 18), 'Occupation'] = 'University Student'
```

---

3. Analyze

3.1 Age Frequency

In the initial analysis of demographics, we can see that 76.5% are aged 19–30 (Figure 1) while the rest are under 19 and up to 69 years old.

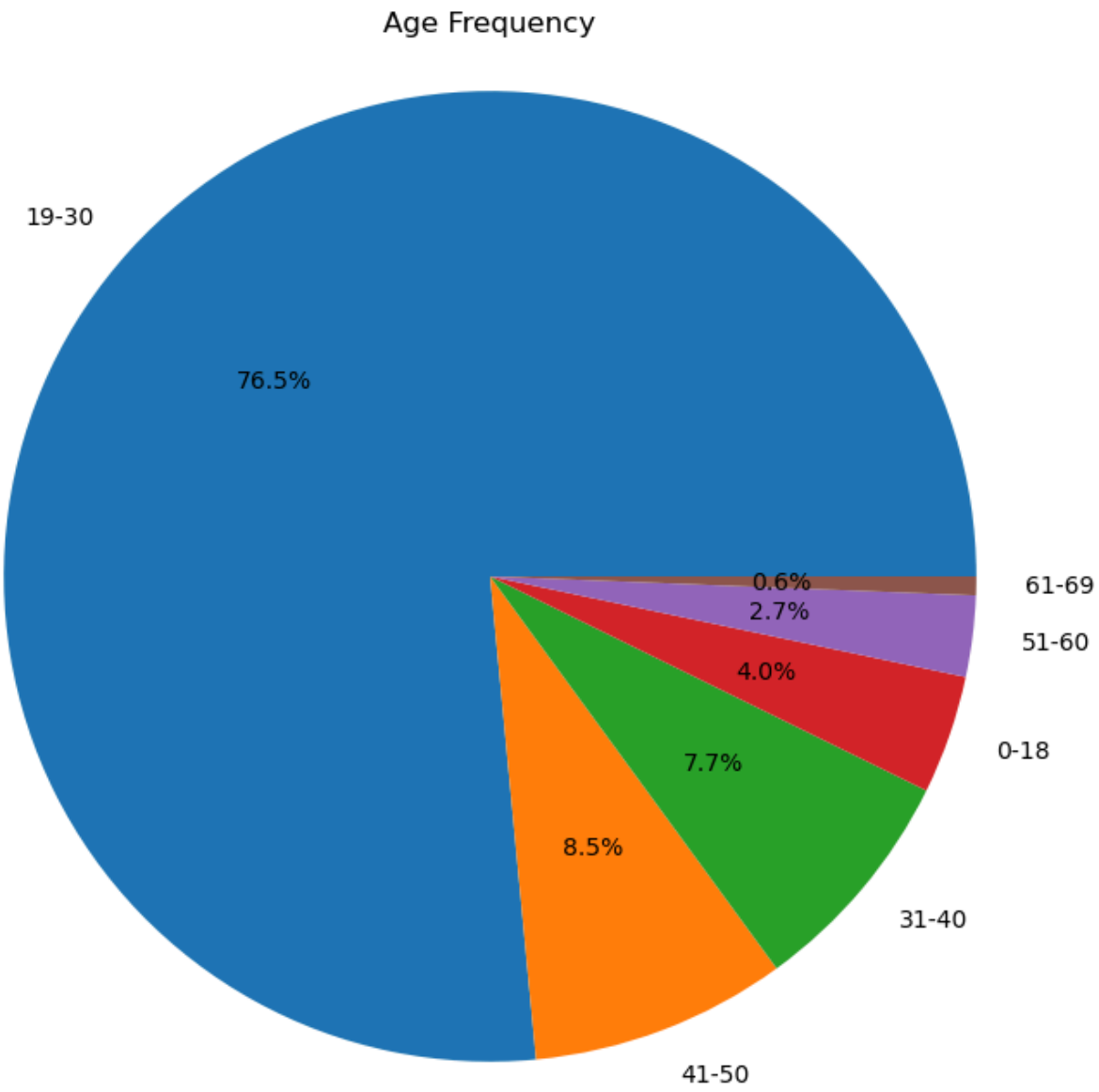


Figure 1: Pie chart of age groups

3.2 Gender Frequency

The majority identify as Female or Male at 54.68% and 47.87% respectively (Figure 2).

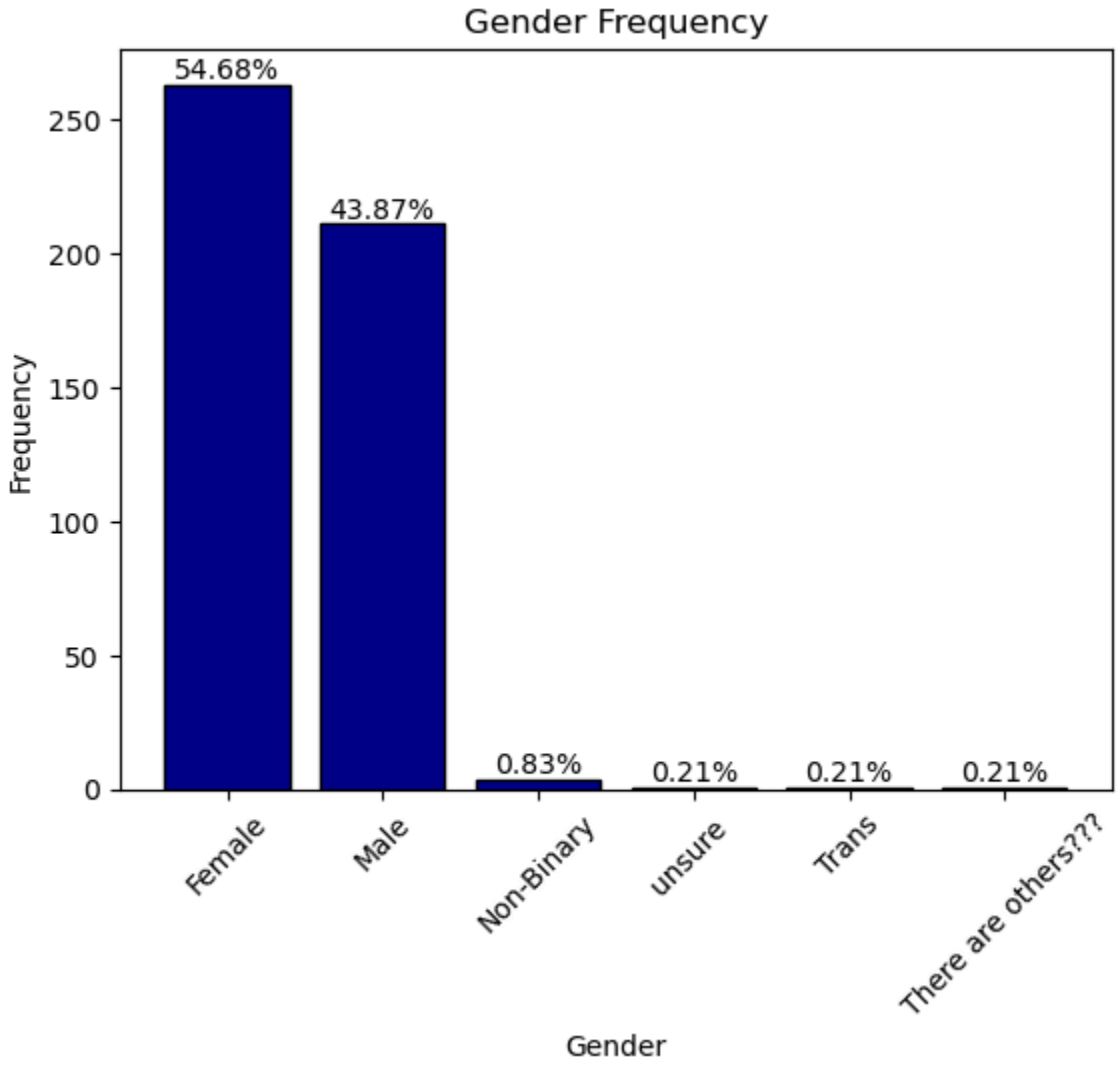


Figure 2: Bar graph of gender frequency

3.3 Relationship Frequency

285 people are Single, 101 are Married, 88 are in a relationship and 7 are Divorced.

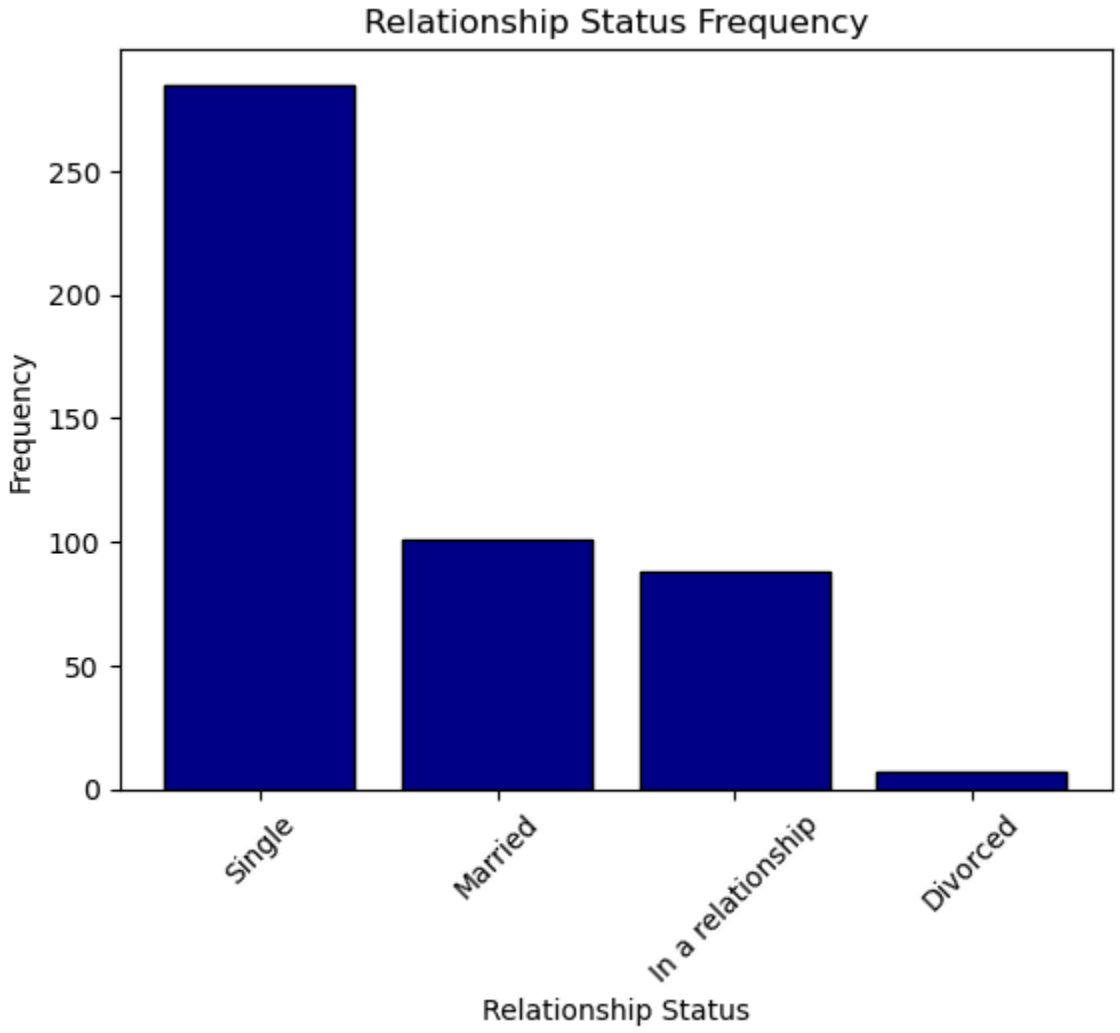


Figure 3: Bar graph of relationship status frequency

3.4 Occupation Frequency

Univeristy Students are the majority at 65.1% with Salaried Workers following behind at 27.4%.

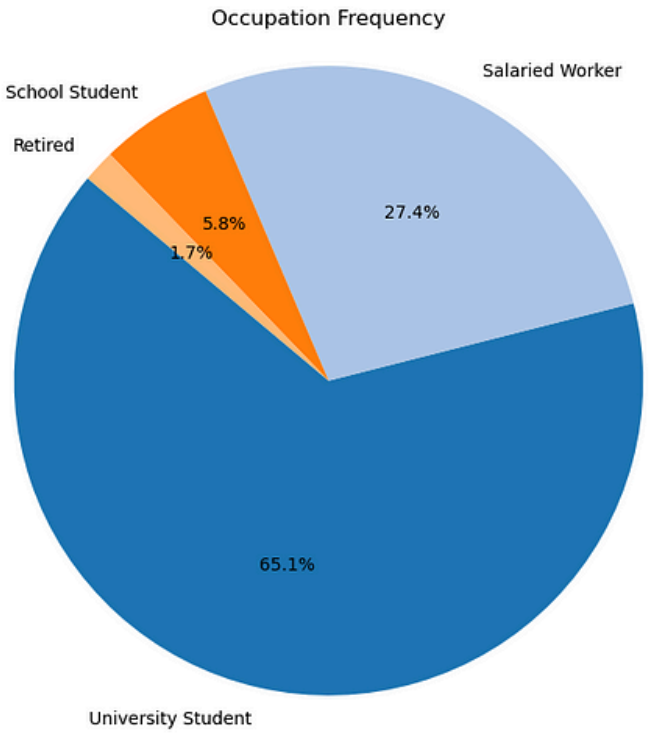


Figure 4: Pie chart of occupation percentages

3.5 Types of Platforms Used Frequency

A combination of Social Networking and Media Sharing is most used at 312 users. 110 used the combination of Social Networking/Media Sharing/Discussion Forum. Social Networking Only is slightly more popular than the remaining platforms and combinations.

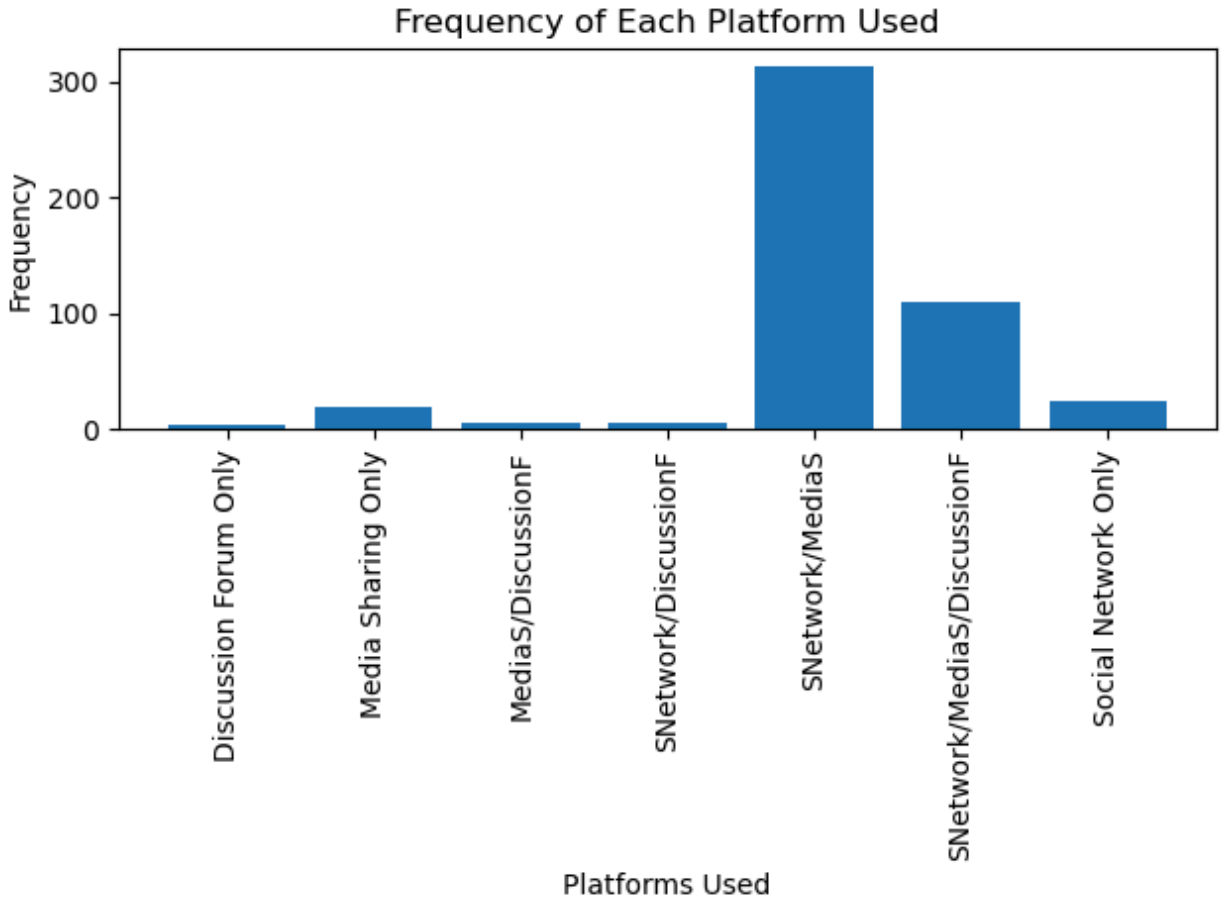


Figure 5: Bar graph of social media platform combinations and their frequencies

3.6 Hours Per Day Frequency

Users were asked how many hours per day did they spend on social media. We can see that there is a gradual rise in the frequency of users as the hours per day interval increases. But, at the 2.5 hours mark, there is a dip in users, less people are using social media at 3.5 hours and 4.5 hours. Then, there is a large jump in users at the 5.5 hours mark.

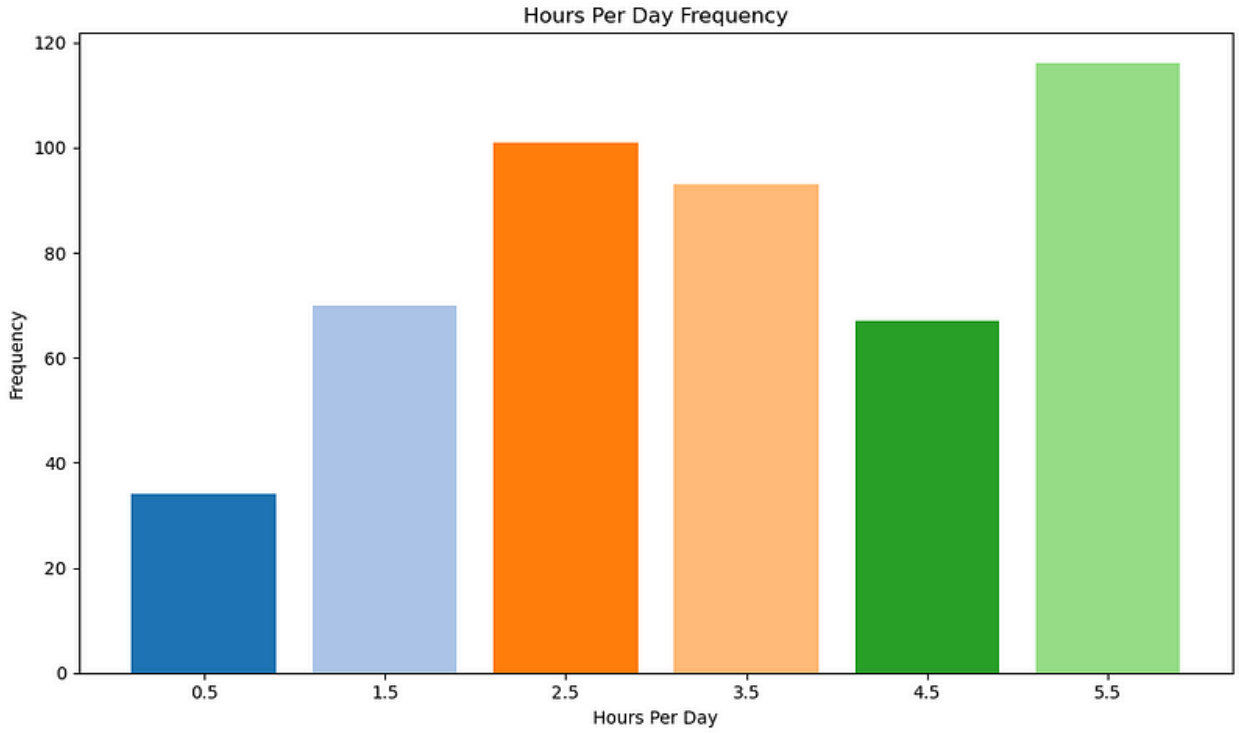


Figure 6: Bar graph showing the frequency of hours per day intervals

3.7 Hours Per Day vs. Platforms Used

We saw in Figure 5 that Social Network/Media sharing has the most users at 312 people as opposed to the 110 people using the triple combination of Social Networking/Media Sharing/Discussion forum.

Here in Figure 7 we see that the triple combination has higher average hours per day use even though it has significantly fewer users.

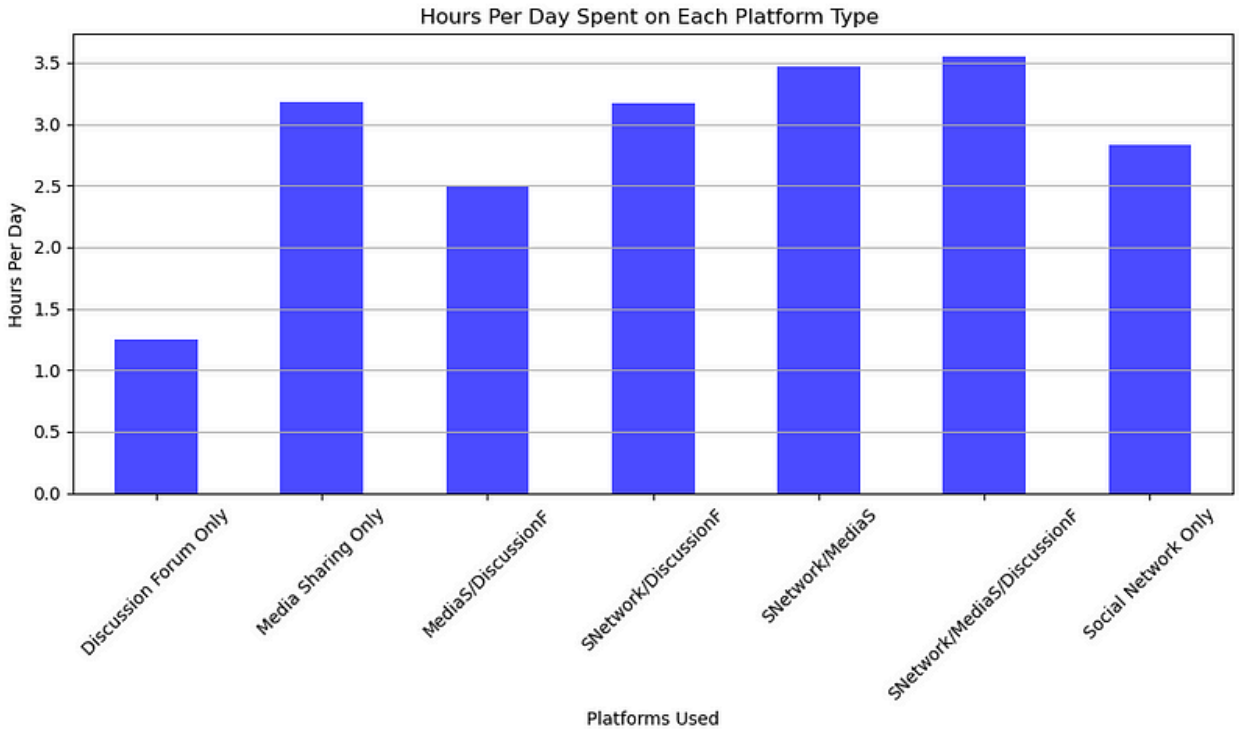
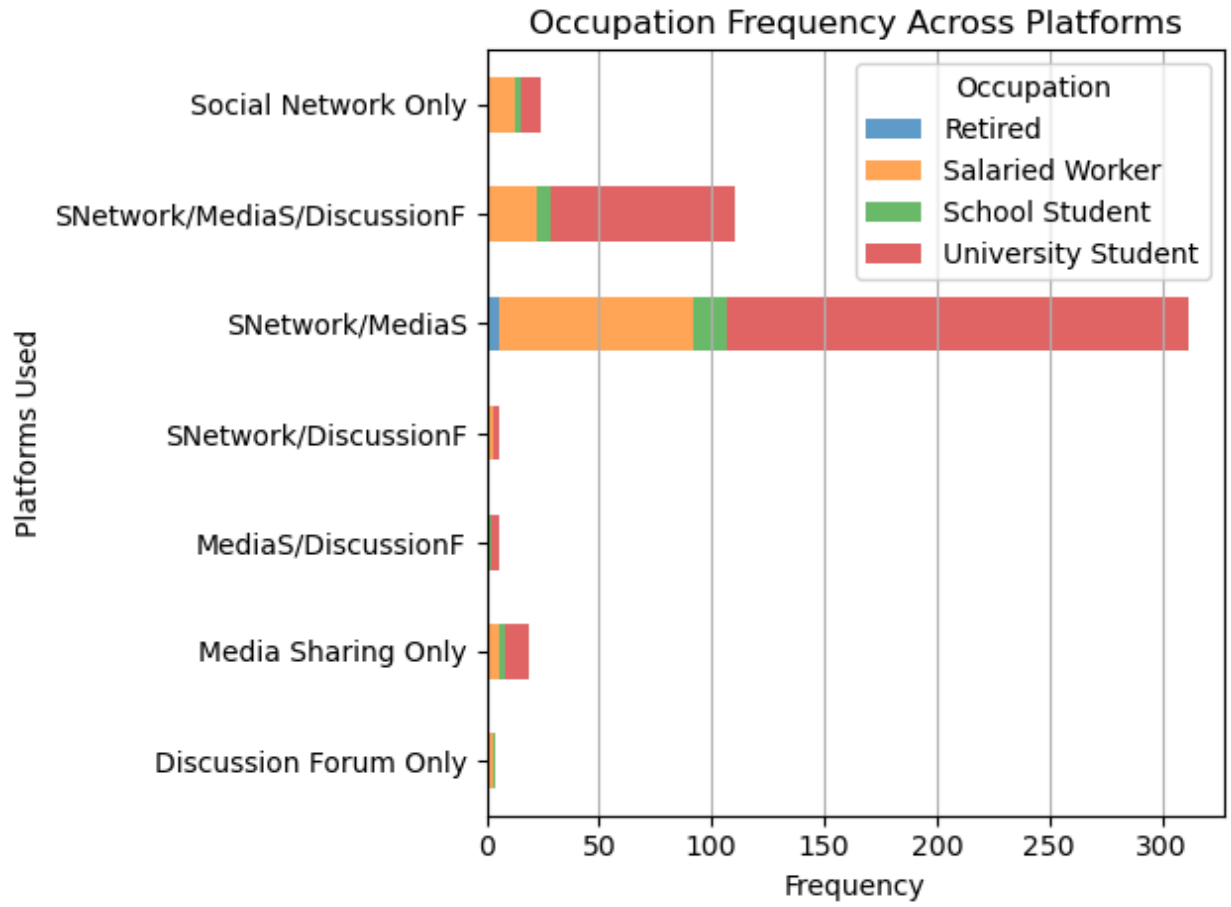


Figure 7: Bar graph showing average hours per day spend on each social media platform

category


Occupation Frequency Across Platforms
Occupation
Retired
Salaried Worker
School Student
University Student
Social Network Only
SNetwork/MediaS/DiscussionF
SNetwork/MediaS
SNetwork/DiscussionF
MediaS/DiscussionF
Media Sharing Only
Discussion Forum Only
Platforms Used
0
50
100
150
200
250
300
Frequency

3.8 Hours Per Day vs. Occupation

School students and University students are spending the most time on social media both over 3.5 hours on average. Retirees come in third with an average of 3.7 hours per day. The Salaried Workers are spending the least amount of time on social media.

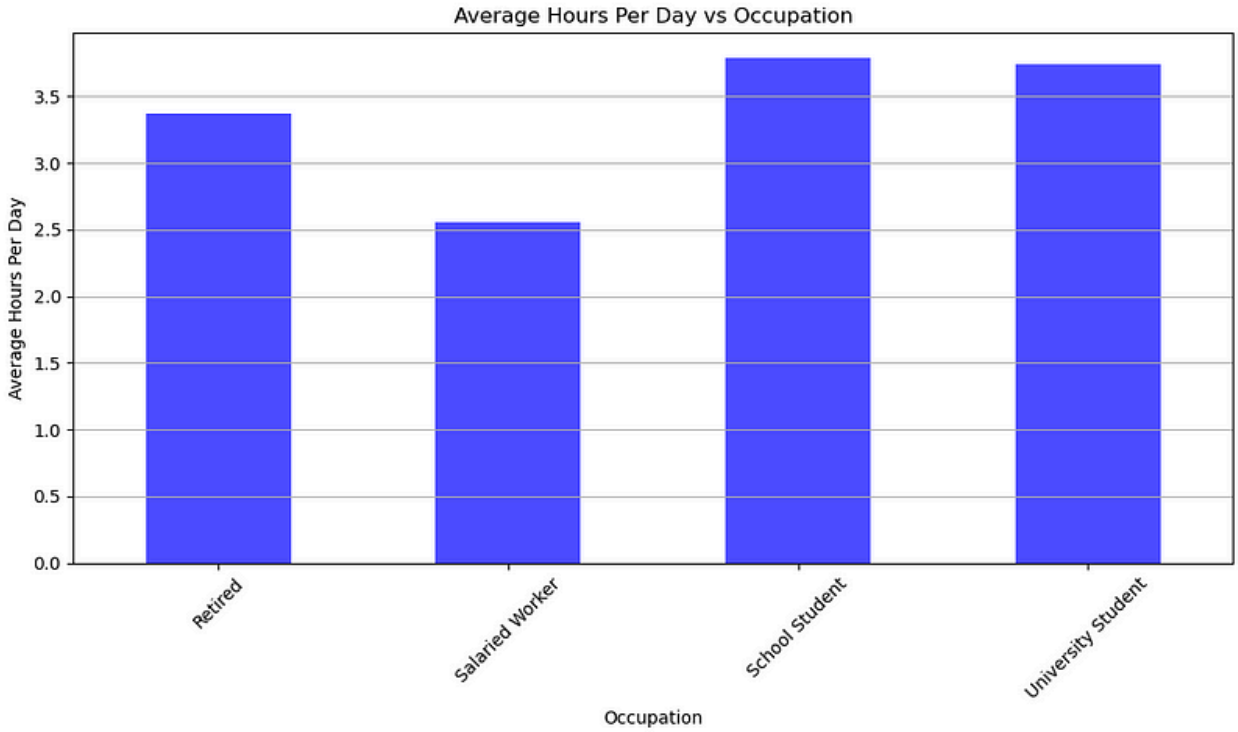


Figure 8: Bar graph showing average hours per day spent on social media for each occupation

3.9 Gender vs Hours Per Day

On average, women spend about 1/2 hour more on social media than men.

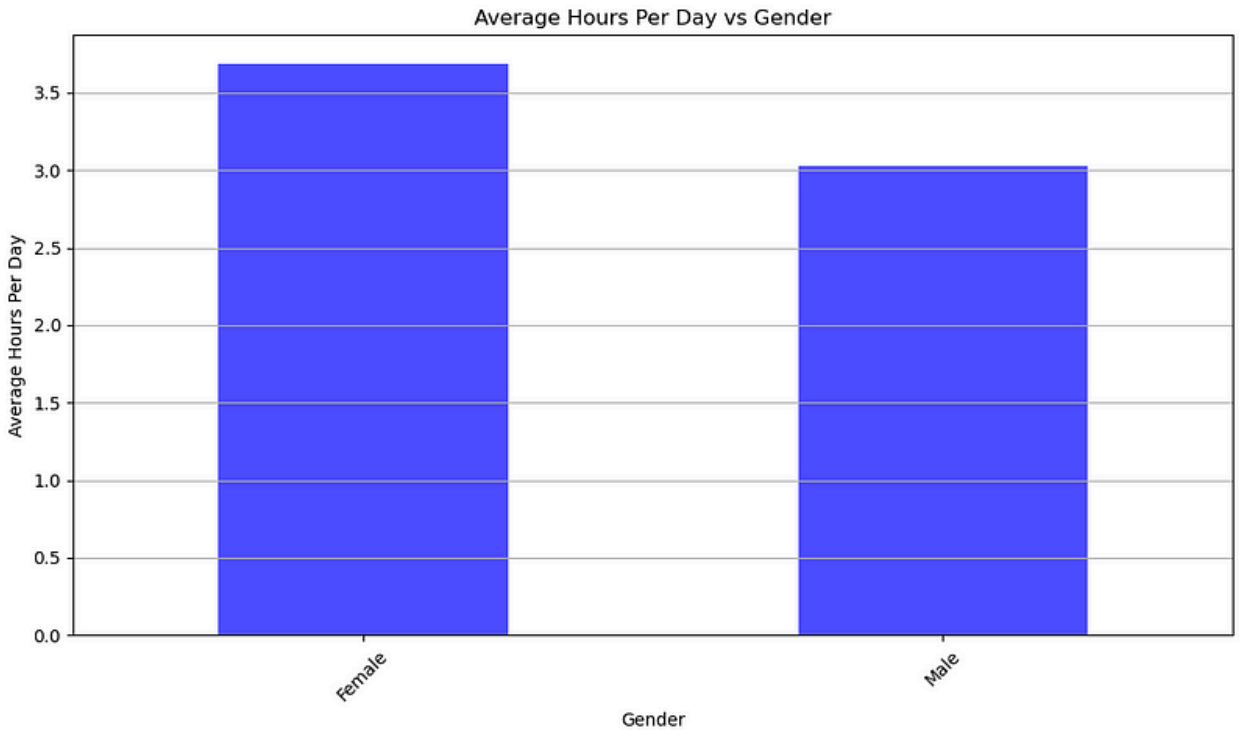


Figure 9: Bar graph showing hours per day spent on social media for each gender.

3.10 Gender vs. Mental Health Screening Score

Women have a slightly higher score than men on average.

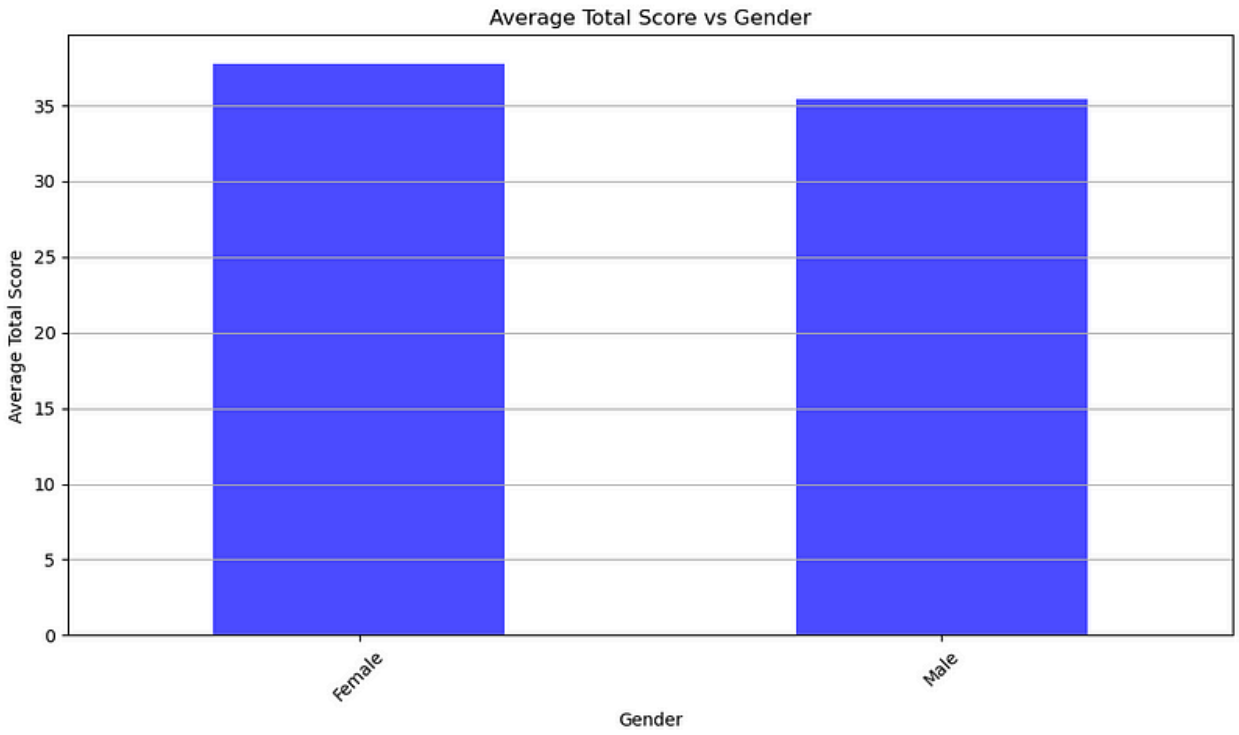


Figure 10: Bar graph showing the average mental health score for each gender.

3.11 Mental Health Screening Scores vs. Age

There is a negative correlation between Age and Total Mental Health Screening Score. Higher scores indicate worse mental health. We can see that as the age increases, scores decrease.

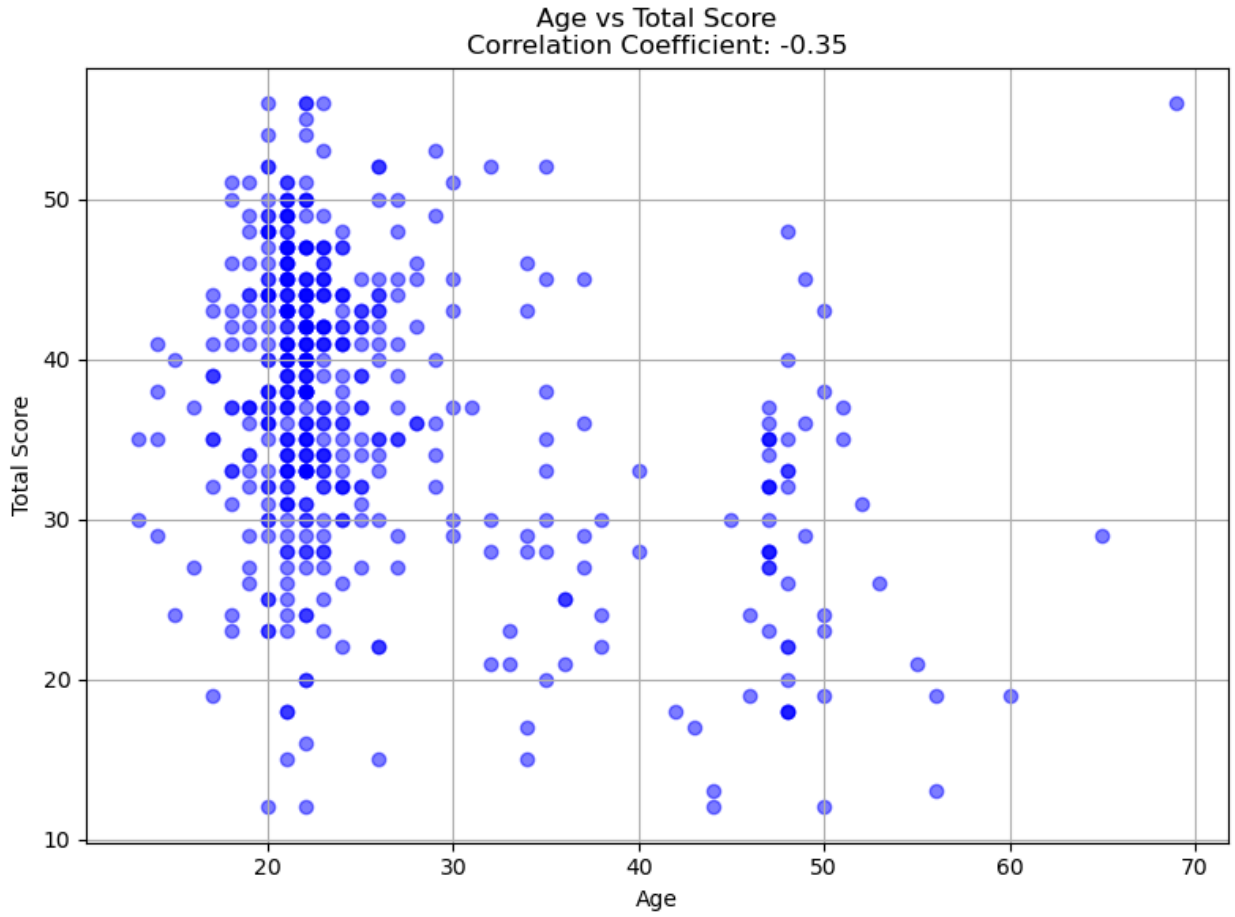


Figure 11: Scatter plot showing the correlation between age and total mental health screening score.

3.12 Mental Health Screening Scores vs. Occupation

University and School students have the highest scores, while salaried workers have the lowest average scores.

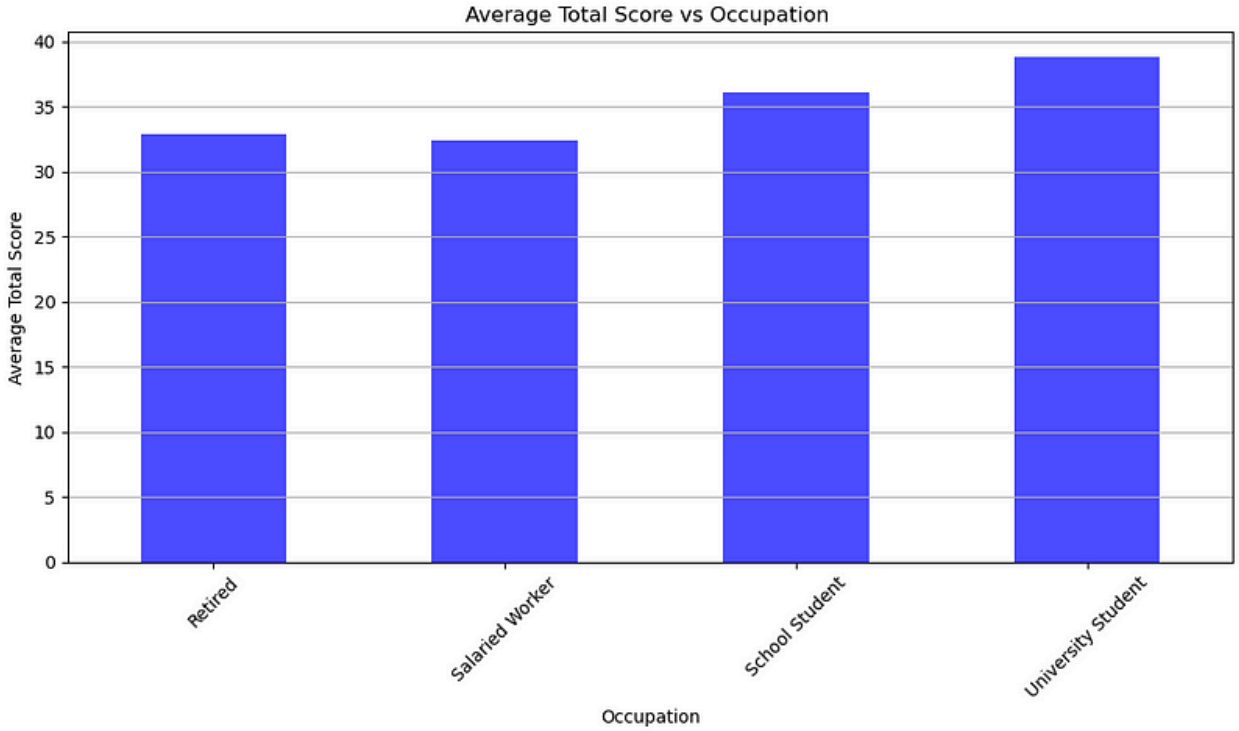


Figure 12: Bar graph showing the average total mental health screening score for each occupation

3.13 Mental Health Screening Scores vs. Hours Per Day

There is a positive correlation between Mental Health Screening Total Score and Hours Per Day spent on social media. We can see that more time spent on social media correlates with higher scores.

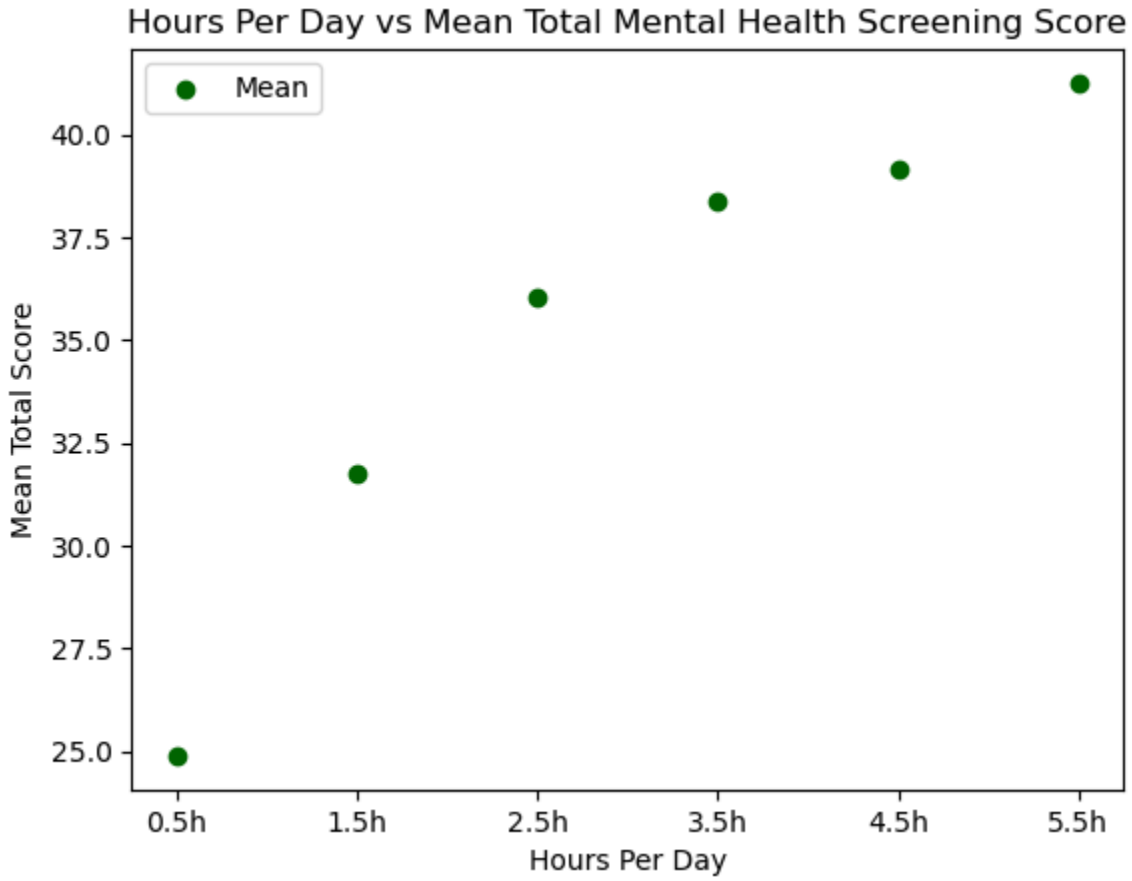


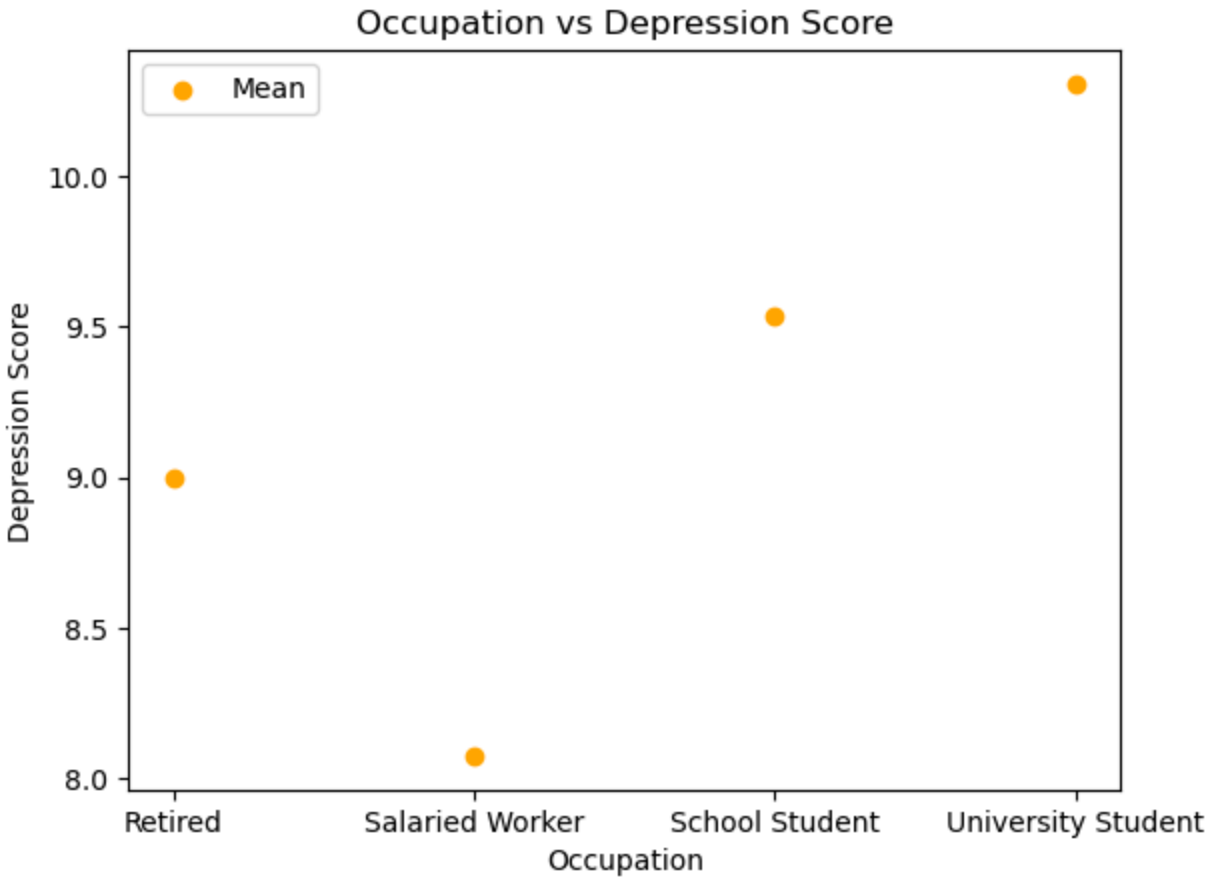
Figure 13: Scatter plot showing the correlation between hours per day spent on social media and average total mental health screening scores.

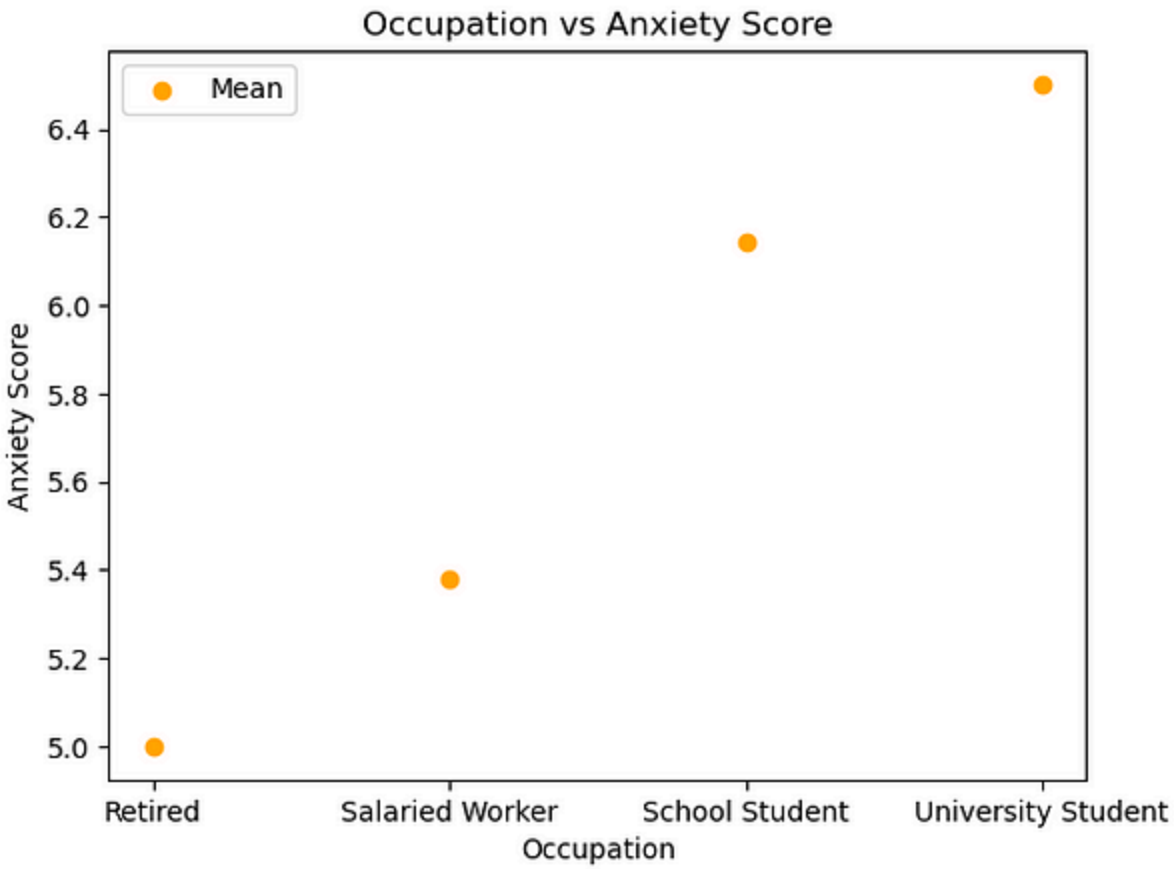
3.14 Mental Health Screening Scores vs. Occupation

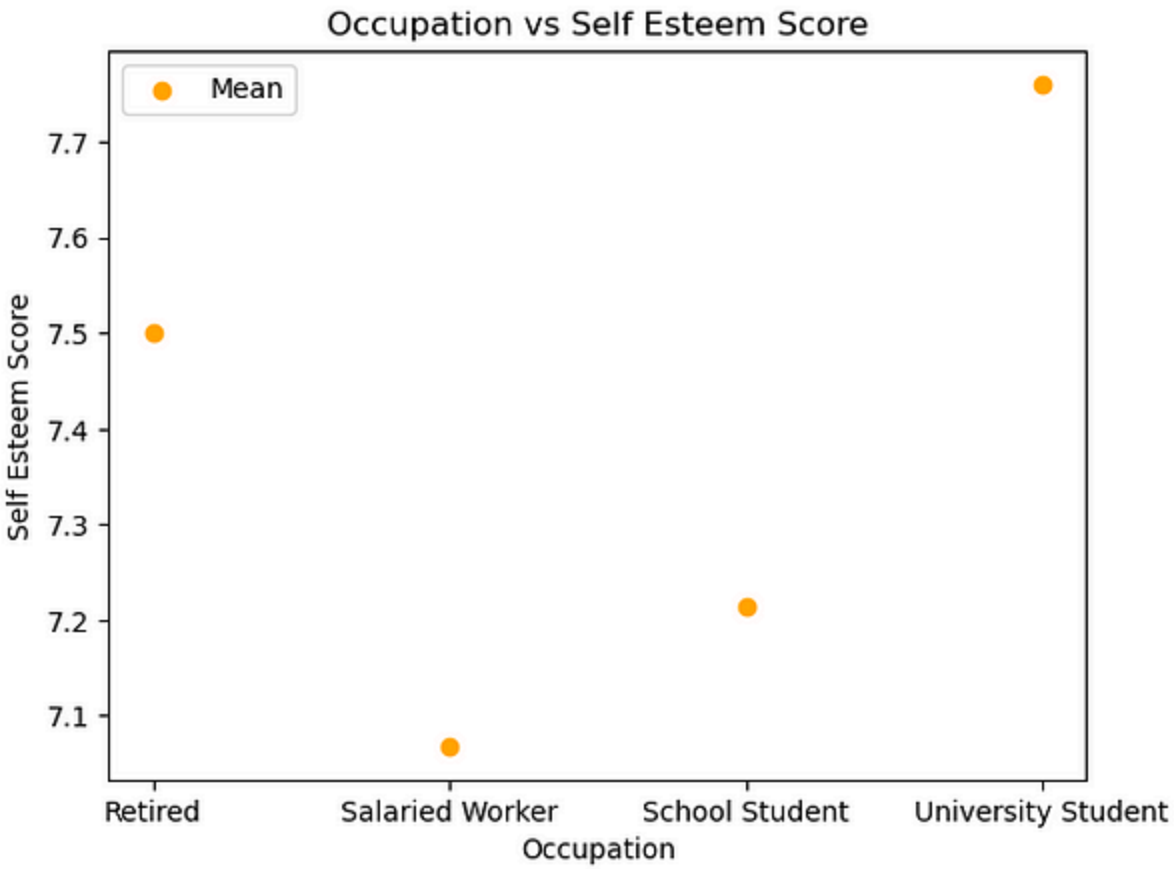
Univeristy students followed by school students have the highest scores across all four screening categories.

Salaried workers have the lowest depression and self-esteem scores. Retirees have higher scores than Salaried Workers.

Retirees have the lowest anxiety scores.


Occupation vs Depression Score
Mean
10.0
9.5
9.0
8.5
8.0
Depression Score
Retired
Salaried Worker
School Student
University Student
Occupation


Occupation vs Anxiety Score
Mean
6.4
6.2
6.0
5.8
5.6
5.4
5.2
5.0
Anxiety Score
Retired
Salaried Worker
School Student
University Student
Occupation


Occupation vs Self Esteem Score
Mean
7.7
7.6
7.5
7.4
7.3
7.2
7.1
Self Esteem Score
Retired
Salaried Worker
School Student
University Student
Occupation

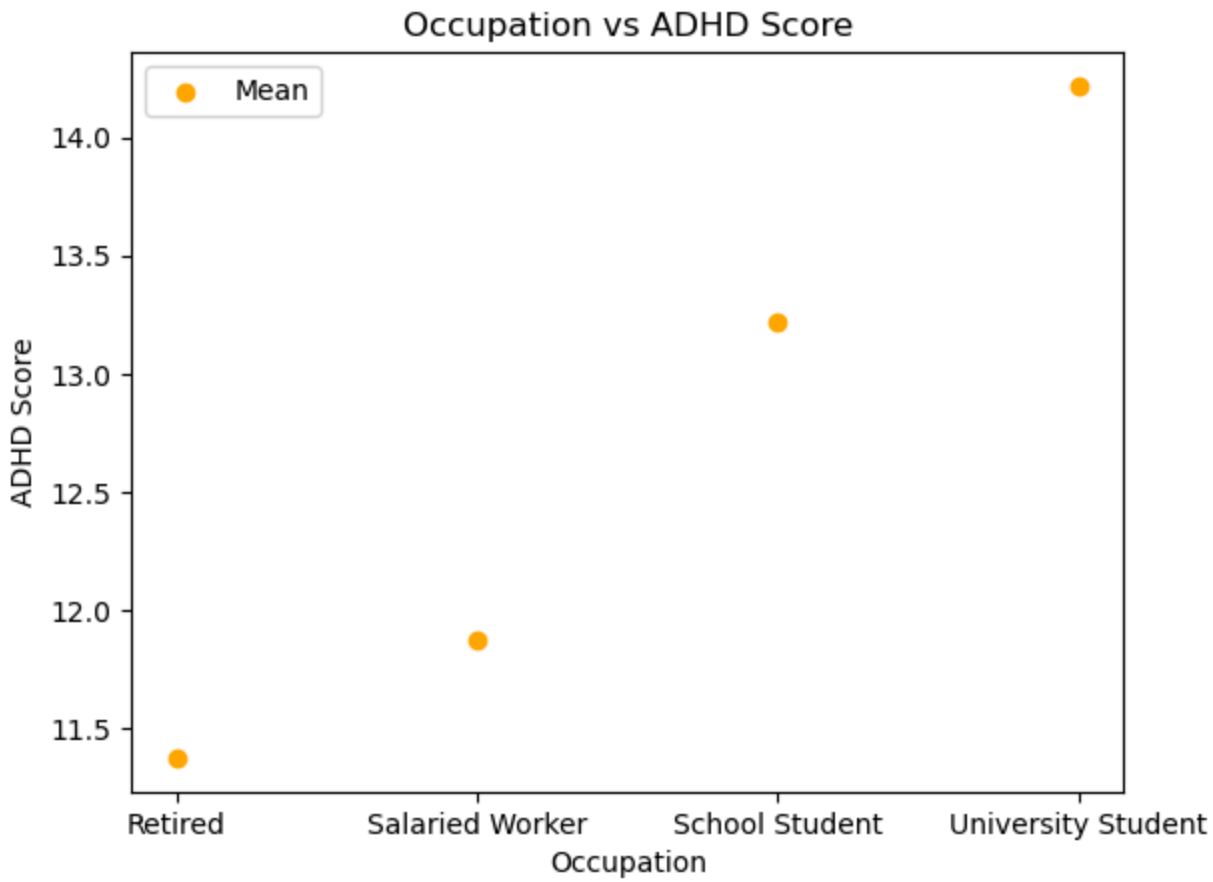


Figure 14: Scatter plots showing average ADHD, Anxiety, Depression, and Self-Esteem scores for each occupation

3.15 Mental Health Screening Total Score vs. Platforms Used

Social Media Platform categories that include Social Networking have higher score but those that include Media Sharing have lower scores.

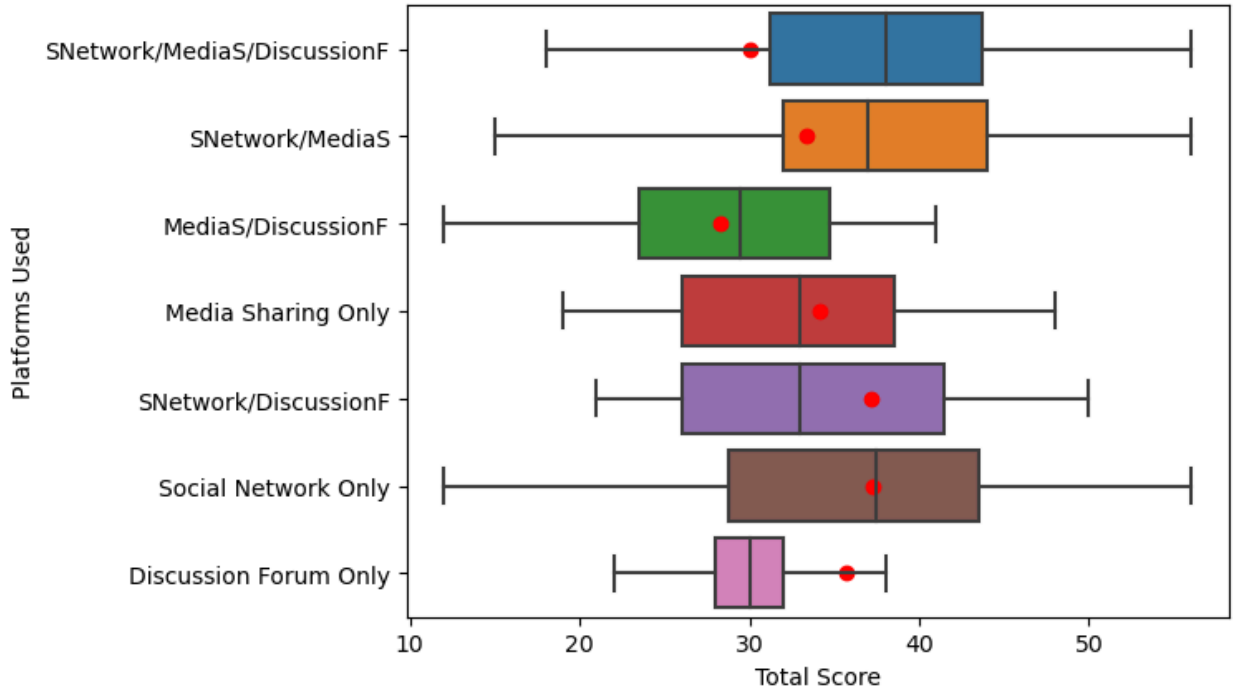


Figure 15: Box plot showing the mean total mental health screening score for each social media platform category.

3.15 Analysis Summary

Univeristy and School students are spending the most time on social media and have higher scores on average in comparison to retirees and salaried workers. Retirees, however, have some higher scores and spend more time on social media than salaried workers.

The frequency of users on social media peaks at 2.5 hours per day, gradually declines through 3.5 and 4.5 hours, and then experiences a sudden spike at in users 5.5 hours.

Social Networking/Media Sharing and Social Networking/Media Sharing/Discussion Forum are the most used social media combinations and the top two in hours per day. Media Sharing Only has significantly less users, but is the third highest hours per day average.

Self-esteem scores are on average lowest in Salaried Workers. Univeristy students have the highest self-esteem score and Retirees second highest.

Social Media Platform categories that include Social Networking have higher score but those that include Media Sharing have lower scores.

4. Share

For every project, the appropriate medium for sharing should considered. In this case, I shared through a report. Analysts can also share by creating a PowerPoint presentation or an interactive dashboard like Tableau.

5. Act

Based on my analysis, I propose the following three recommendations.

1. Encourage users to limit their daily social media consumption to a healthy range. Educational campaigns or features within social media platforms could provide users with insights into their usage habits and encourage mindful consumption.
2. Push in-app messages that encourage positive interaction to improve Self-esteem.
3. Define the purpose or objective of your platform in a way that encourages users to engage meaningfully rather than as a mere distraction.
4. Collect further data to gather insights on user engagement patterns, specifically focusing on the days and times when users are most active on social media platforms to assess potential interference with daily activities or sleep patterns.
