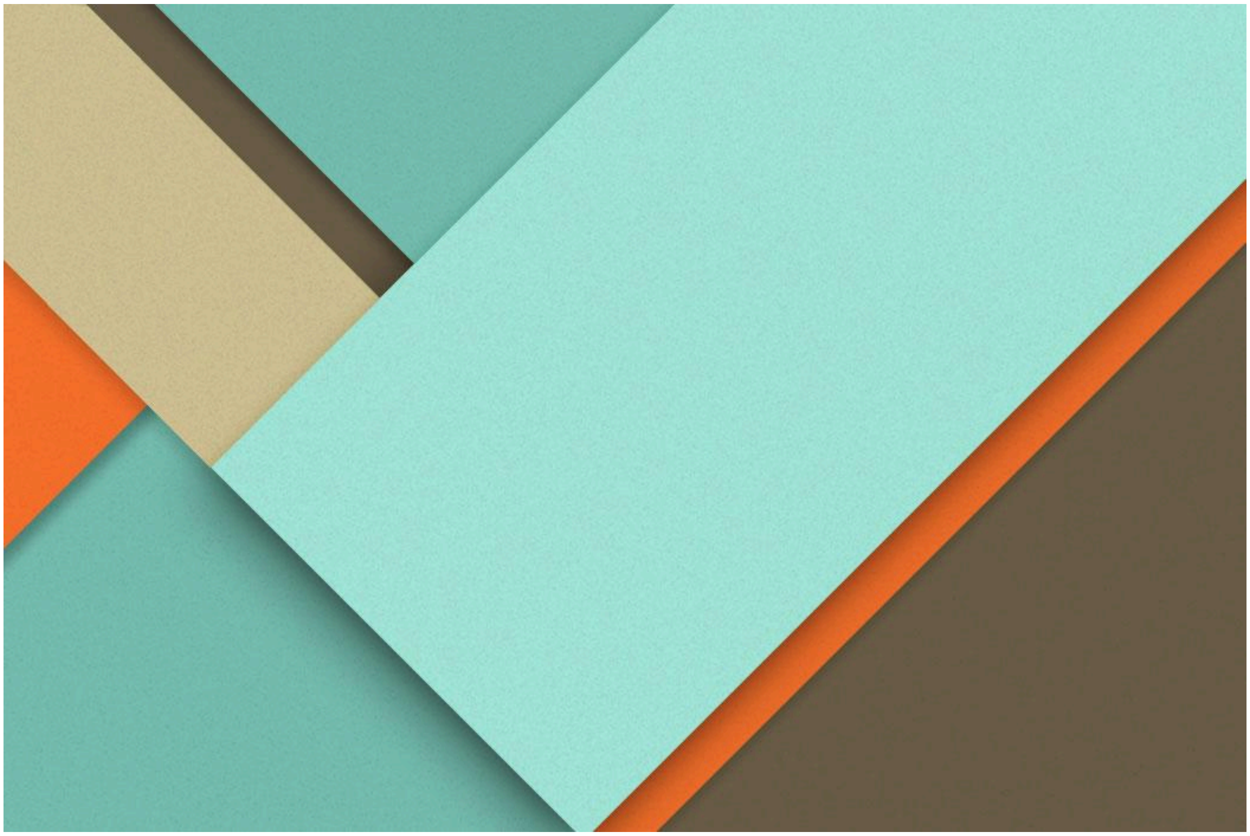


Exploring the Relationship Between Social Media Usage and Anxiety

25.01.2026

Hayley Lee

Overview

For this project, a dataset containing individuals' social media usage patterns and clinically reported anxiety levels will be used to support a machine learning classification task. The objective is to train and evaluate a model that can predict a user's anxiety severity based on features derived from social media behavior.

Goals

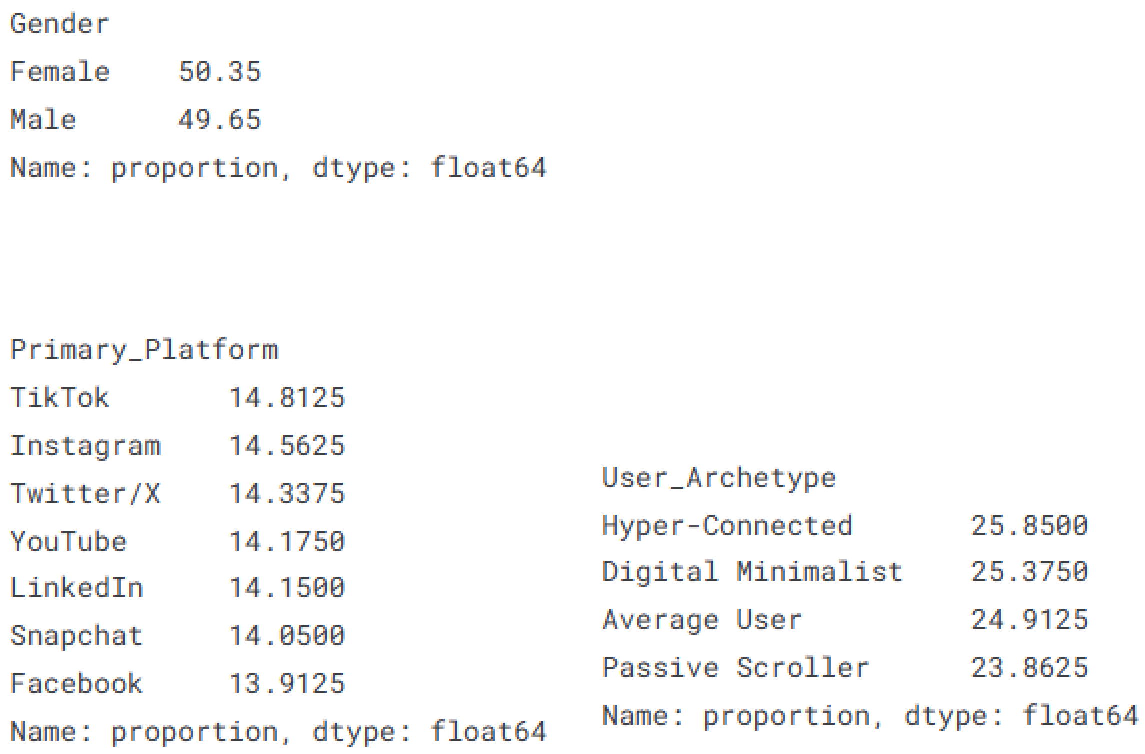
1. Select a suitable dataset for the above task
2. Check dataset quality and clean (if needed)
3. Conduct data exploration using visualisations

Dataset Selection

The following dataset from kaggle was chosen for this project: https://www.kaggle.com/datasets/bertnardomariouskono/social-media-and-mental-health

This dataset captures the relationship between social media use and mental well-being through a synthetic but scientifically grounded log of user behavior and psychological outcomes. It contains 8,000 entries combining social media usage patterns with standardized anxiety and depression measures, including gender to support comparative and statistical analysis. Key features cover usage intensity, psychometric scores (GAD-7 and PHQ-9), and behavioral risk factors such as social comparison and sleep displacement.

While exploring key features such as user gender, primary social media platform, and archetypes (e.g. hyper-connected, digital minimalist, etc), the dataset was found to be well balanced. Figure 1 shows the percentage split of data belonging to certain features.



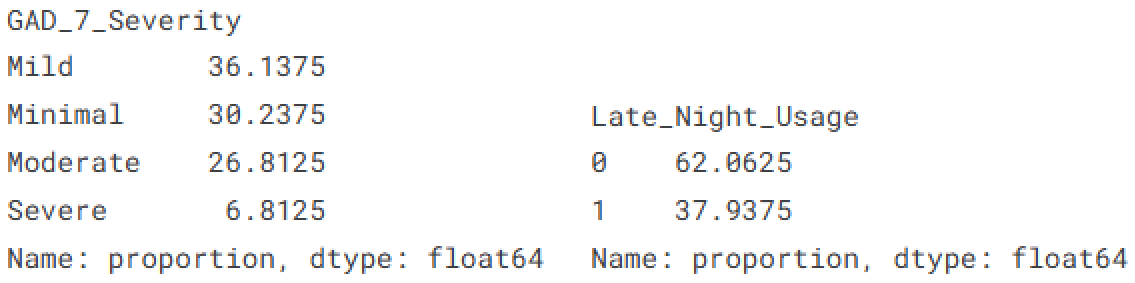
```
Gender
Female    50.35
Male      49.65
Name: proportion, dtype: float64
```

```
Primary_Platform
TikTok       14.8125
Instagram    14.5625
Twitter/X    14.3375
YouTube      14.1750
LinkedIn     14.1500
Snapchat     14.0500
Facebook     13.9125
Name: proportion, dtype: float64
```

```
User_Archetype
Hyper-Connected      25.8500
Digital Minimalist   25.3750
Average User         24.9125
Passive Scroller     23.8625
Name: proportion, dtype: float64
```

Figure 1: The percentage split of data for some feature categories.

Dataset limitations could be introduced by features such as Late_Night_Usage and GAD_7_Severity, as these were found to possess class imbalance between their possible values (see Figure 2).



```
GAD_7_Severity
Mild        36.1375
Minimal     30.2375
Moderate    26.8125
Severe       6.8125
Name: proportion, dtype: float64
```

```
Late_Night_Usage
0    62.0625
1    37.9375
Name: proportion, dtype: float64
```

Figure 2: The percentage split of data for key feature categories. There is a lack of data for severe anxiety cases and reported late night usage, which may affect reliability of results.

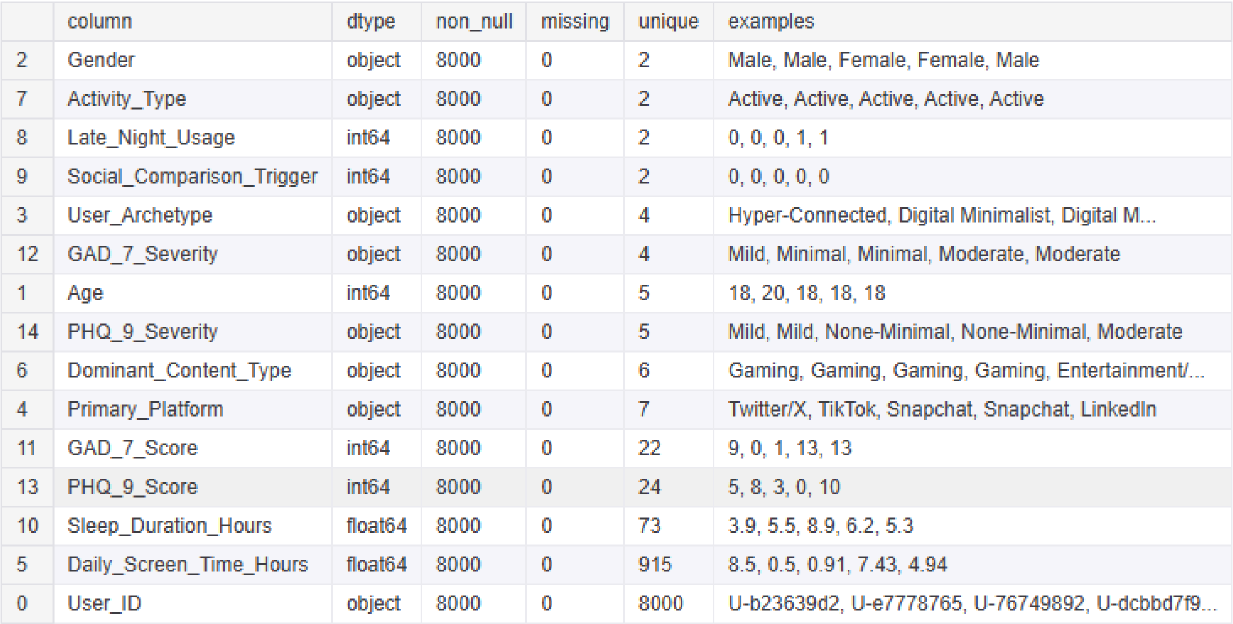
Data Preprocessing

A data glossary was created to complete the following goals:

- Identify missing values

- Identify unique values (determines whether values are distinct or continuous)
- Show dataset size
- Identify columns present in dataset

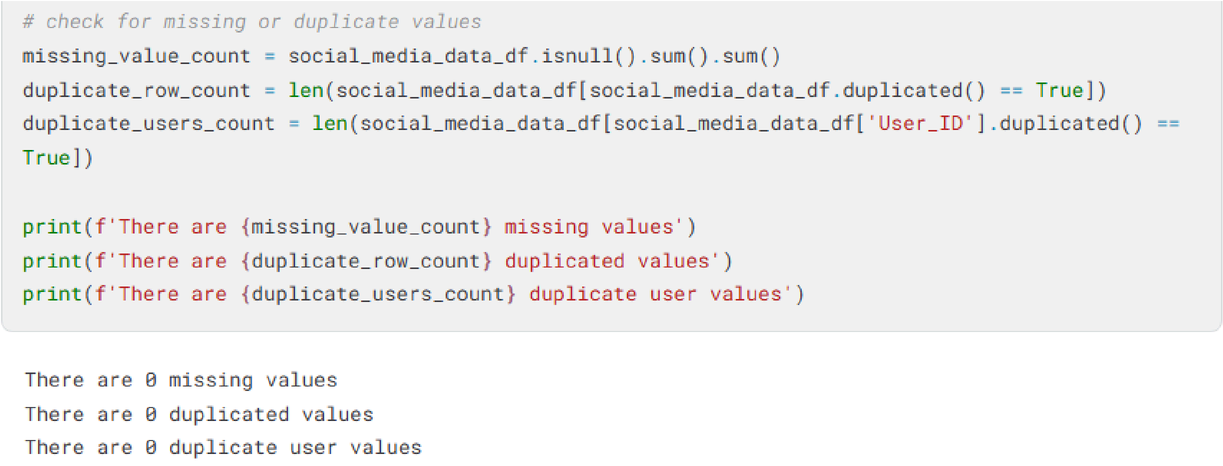
Table 1 shows the glossary output:



	column	dtype	non_null	missing	unique	examples
2	Gender	object	8000	0	2	Male, Male, Female, Female, Male
7	Activity_Type	object	8000	0	2	Active, Active, Active, Active, Active
8	Late_Night_Usage	int64	8000	0	2	0, 0, 0, 1, 1
9	Social_Comparison_Trigger	int64	8000	0	2	0, 0, 0, 0, 0
3	User_Archetype	object	8000	0	4	Hyper-Connected, Digital Minimalist, Digital M...
12	GAD_7_Severity	object	8000	0	4	Mild, Minimal, Minimal, Moderate, Moderate
1	Age	int64	8000	0	5	18, 20, 18, 18, 18
14	PHQ_9_Severity	object	8000	0	5	Mild, Mild, None-Minimal, None-Minimal, Moderate
6	Dominant_Content_Type	object	8000	0	6	Gaming, Gaming, Gaming, Gaming, Entertainment/...
4	Primary_Platform	object	8000	0	7	Twitter/X, TikTok, Snapchat, Snapchat, LinkedIn
11	GAD_7_Score	int64	8000	0	22	9, 0, 1, 13, 13
13	PHQ_9_Score	int64	8000	0	24	5, 8, 3, 0, 10
10	Sleep_Duration_Hours	float64	8000	0	73	3.9, 5.5, 8.9, 6.2, 5.3
5	Daily_Screen_Time_Hours	float64	8000	0	915	8.5, 0.5, 0.91, 7.43, 4.94
0	User_ID	object	8000	0	8000	U-b23639d2, U-e7778765, U-76749892, U-dcbbd7f9...

Table 1: Data glossary showing key metrics such as size of dataset, number of missing values, columns present, etc.

Further checks were created to test for duplicate user and row entries in the data (Figure 3).



```
# check for missing or duplicate values
missing_value_count = social_media_data_df.isnull().sum().sum()
duplicate_row_count = len(social_media_data_df[social_media_data_df.duplicated() == True])
duplicate_users_count = len(social_media_data_df[social_media_data_df['User_ID'].duplicated() ==
True])

print(f'There are {missing_value_count} missing values')
print(f'There are {duplicate_row_count} duplicated values')
print(f'There are {duplicate_users_count} duplicate user values')
```

```
There are 0 missing values
There are 0 duplicated values
There are 0 duplicate user values
```

Figure 3: Checks for missing values, duplicated rows or duplicate user entries in the dataset.

Since the dataset is synthetically generated, it contains no missing values or duplicate records. If such issues were present, they would be addressed as follows:

- Missing values: Values such as sleep duration and daily screen time could be handled by taking an average of known values from users with similar features. Missing values from categorical data such as gender could be replaced with a new 'Unknown' category. Missing GAD-7 or PHQ-9 severities could be approximated using the average GAD/PHQ scores of users with the same severity category.
- Duplicate values: deduplication of rows in the dataframe can be achieved through pandas easily. Removal of rows from the same user would keep the dataset balanced.

Data anonymisation was not needed as the usernames are synthetically created and there is no data which can identify a real person. If this was required, a randomised string could be produced to replace all social media user names in the dataset.

Data Exploration and Visualisation

The following visualisations examine potential variables that may affect the relationship between social media use and individual anxiety levels.

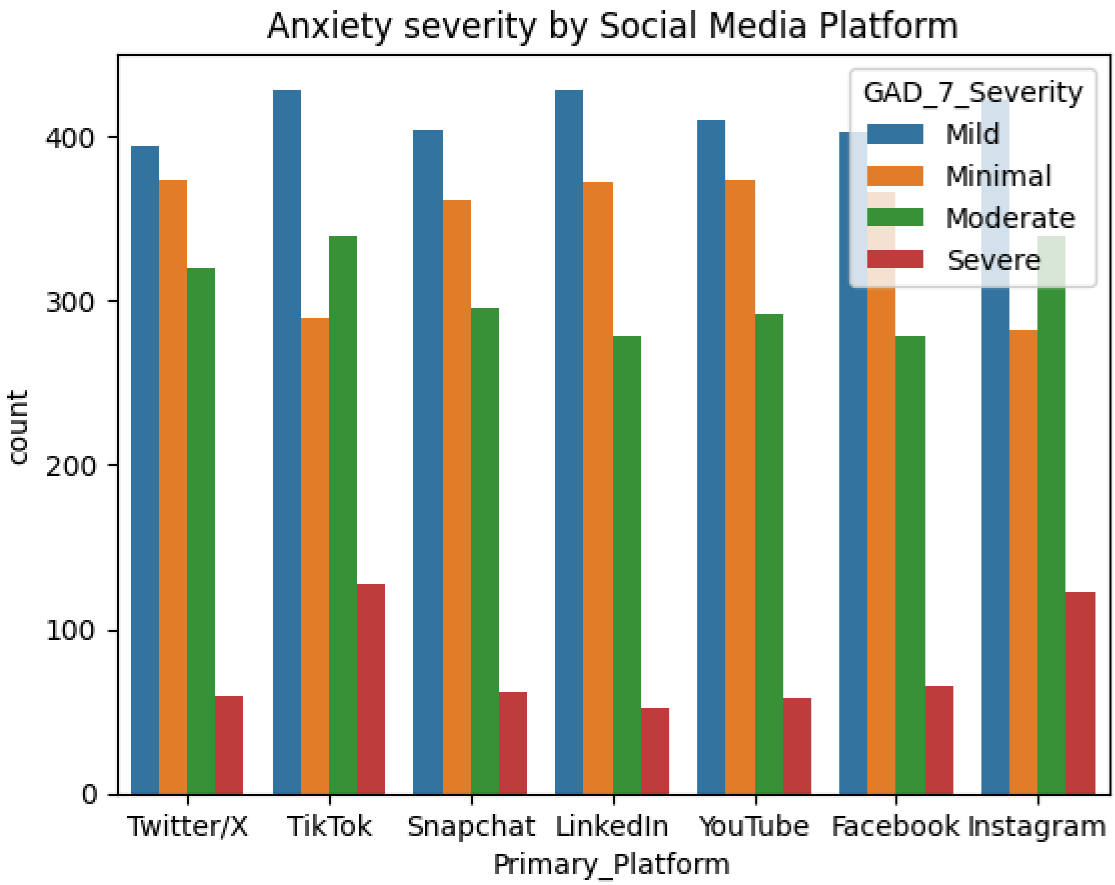


Figure 4: A bar chart showing the count of users within specific anxiety severity categories and their primary social media platform.

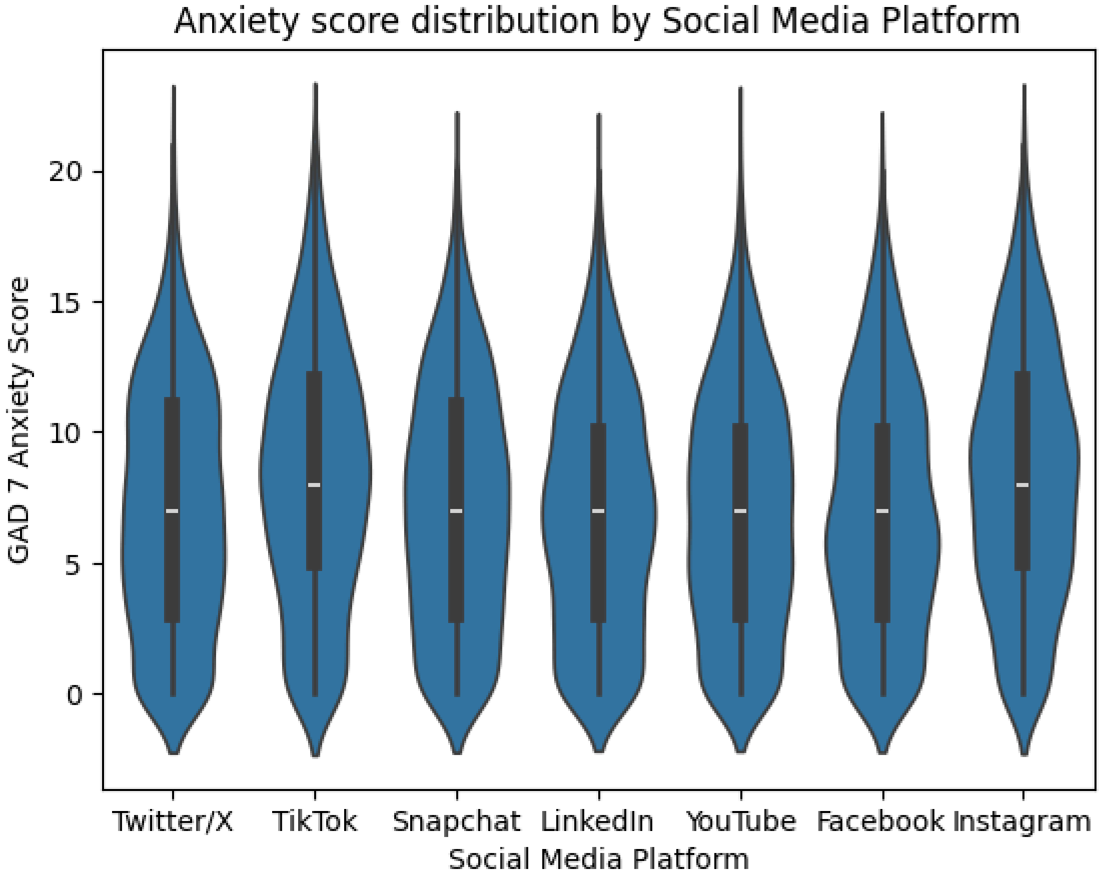


Figure 5: A violin plot showing the range and distribution of anxiety scores for each social media platform's users.

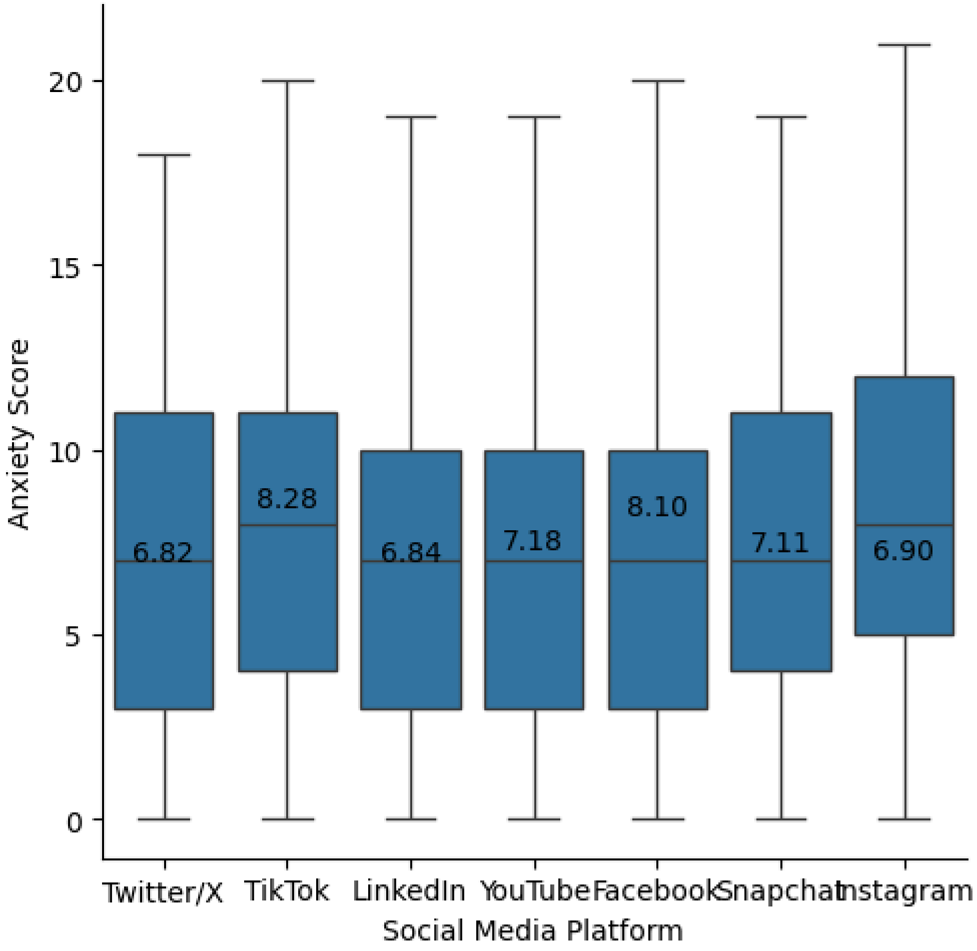


Figure 6: A catplot showing the range of male user anxiety scores per their main social media platform.

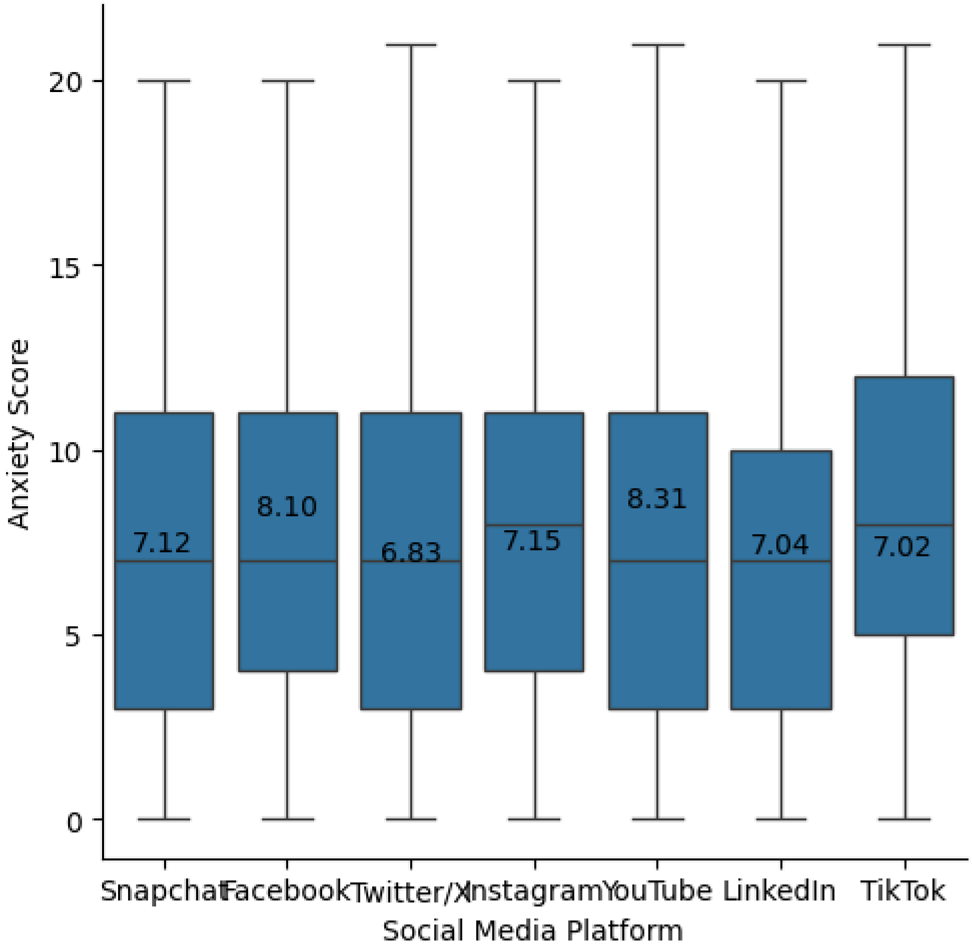


Figure 7: A catplot showing the range of female user anxiety scores per their main social media platform.

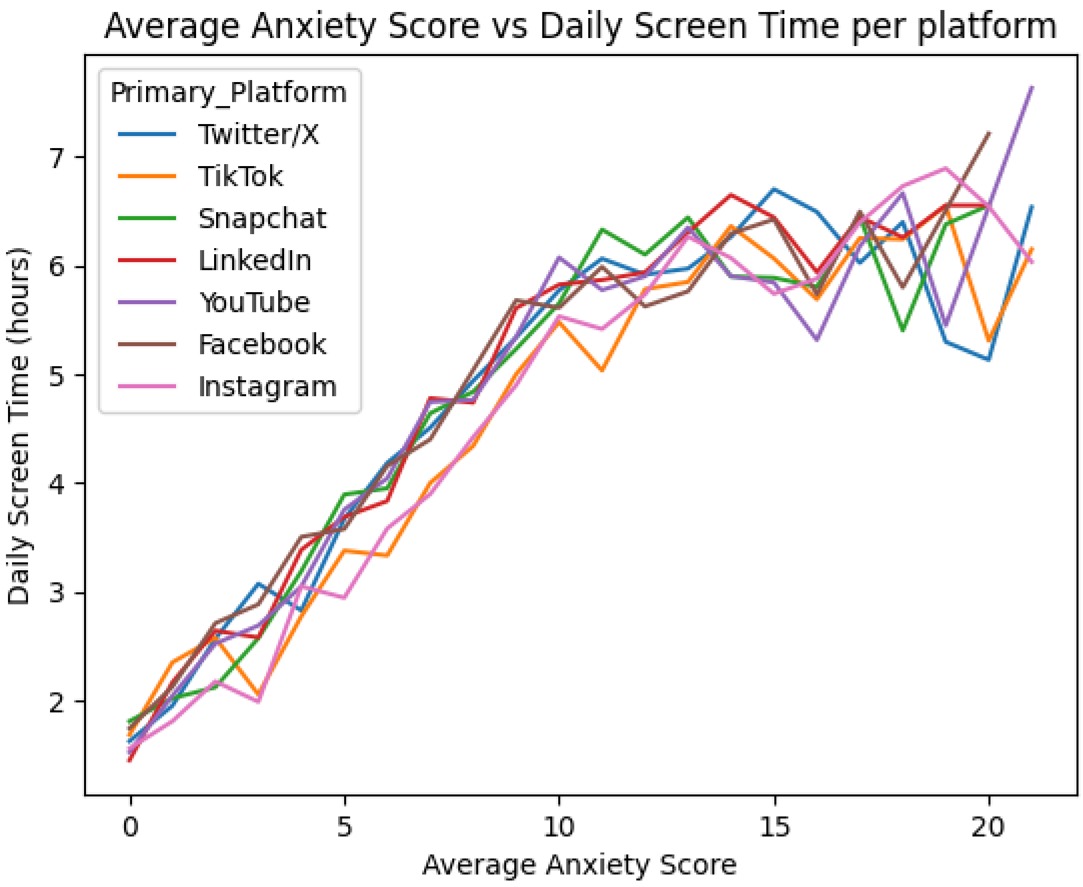


Figure 8: A graph showing the relationship between anxiety score and daily screen time, per social media platform.

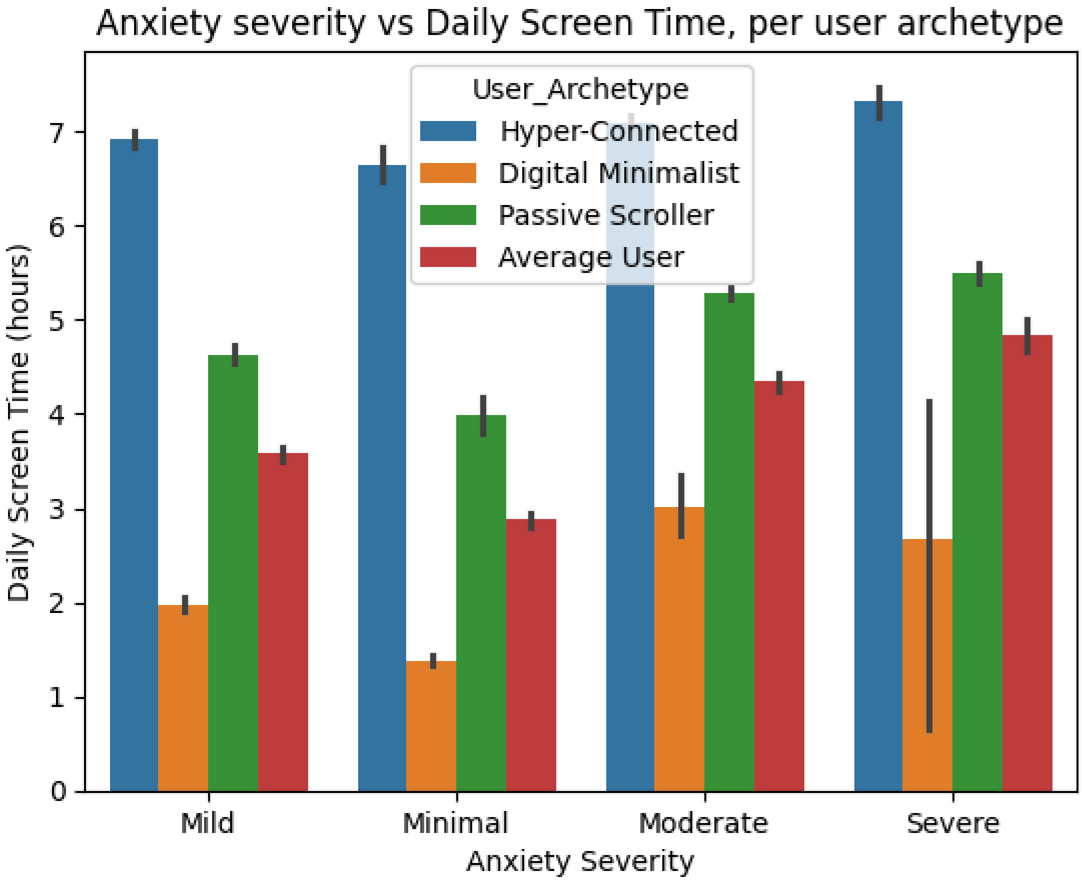


Figure 9: Barchart showing the daily screen time of each user archetype, with their associated anxiety severity category.

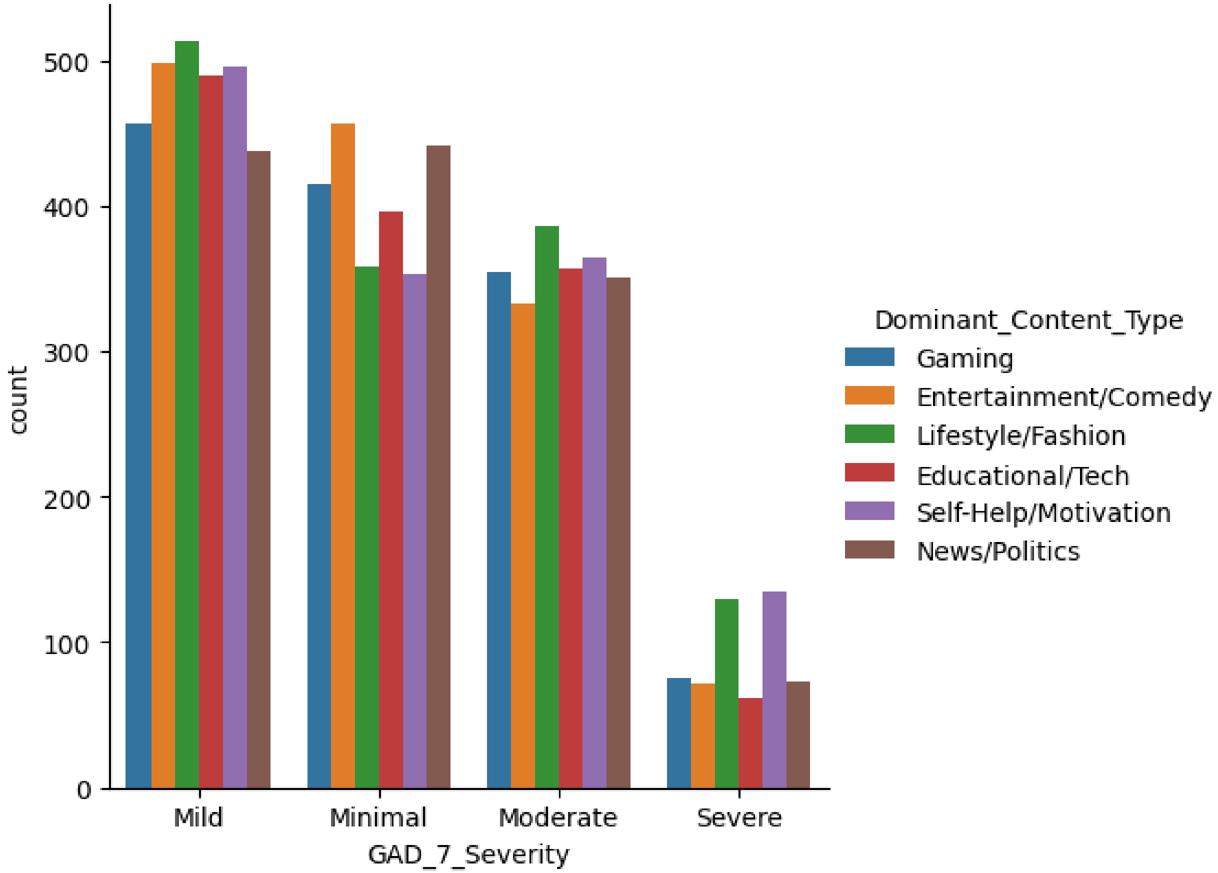


Figure 10: Barchart showing the count of users in each anxiety severity category and their dominant content type.

Conclusions based on visualisation results:

- Figure 4 shows that Tiktok and Instagram have the highest number of users that fall under severe or moderate severities for anxiety.
- Figure 6 and 7 show that compared to male users, women have higher mean anxiety scores on Instagram, Youtube and LinkedIn, while men have higher average anxiety scores compared to female users on Tiktok. Both genders had higher average anxiety scores on both Youtube and Facebook.
- Figure 8 shows a positive correlation between a user’s anxiety score and their daily screen time.
- Figure 10 shows an even spread of dominant content type amongst users in different anxiety severity classes, except for severe. The most popular categories for severe seem to be lifestyle/fashion and self help/motivation.

From the above diagrams, the amount of daily screen time, the gender of the user, their dominant social media platform and their preferred content type will be key features used to determine a user's anxiety severity.