In [1]:
```
import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
```

In [2]:
```
df = pd.read_csv(r"D:\data analysis\children addiction\Students Social Media
df
```

Out[2]:

	Student_ID	Age	Gender	Academic_Level	Country	Avg_Daily_Usage_I
0	1	19	Female	Undergraduate	Bangladesh	
1	2	22	Male	Graduate	India	
2	3	20	Female	Undergraduate	USA	
3	4	18	Male	High School	UK	
4	5	21	Male	Graduate	Canada	
...	...	...	...	...	...	
700	701	20	Female	Undergraduate	Italy	
701	702	23	Male	Graduate	Russia	
702	703	21	Female	Undergraduate	China	
703	704	24	Male	Graduate	Japan	
704	705	19	Female	Undergraduate	Poland	

705 rows × 13 columns

In [3]:
```
df.describe()
```

Out[3]:

	Student_ID	Age	Avg_Daily_Usage_Hours	Sleep_Hours_Per_Night
count	705.000000	705.000000	705.000000	705.000000
mean	353.000000	20.659574	4.918723	6.868936
std	203.660256	1.399217	1.257395	1.126848
min	1.000000	18.000000	1.500000	3.800000
25%	177.000000	19.000000	4.100000	6.000000
50%	353.000000	21.000000	4.800000	6.900000
75%	529.000000	22.000000	5.800000	7.700000
max	705.000000	24.000000	8.500000	9.600000

In [4]:
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 705 entries, 0 to 704
Data columns (total 13 columns):
 #   Column                          Non-Null Count  Dtype  
---  ------                          --------------  -----  
 0   Student_ID                      705 non-null    int64  
 1   Age                             705 non-null    int64  
 2   Gender                          705 non-null    object 
 3   Academic_Level                  705 non-null    object 
 4   Country                         705 non-null    object 
 5   Avg_Daily_Usage_Hours           705 non-null    float64
 6   Most_Used_Platform              705 non-null    object 
 7   Affects_Academic_Performance    705 non-null    object 
 8   Sleep_Hours_Per_Night           705 non-null    float64
 9   Mental_Health_Score             705 non-null    int64  
 10  Relationship_Status             705 non-null    object 
 11  Conflicts_Over_Social_Media     705 non-null    int64  
 12  Addicted_Score                  705 non-null    int64  
dtypes: float64(2), int64(5), object(6)
memory usage: 71.7+ KB
```

In [5]:
```
df.isnull().sum()   #Give the sum of the null values in each column
```

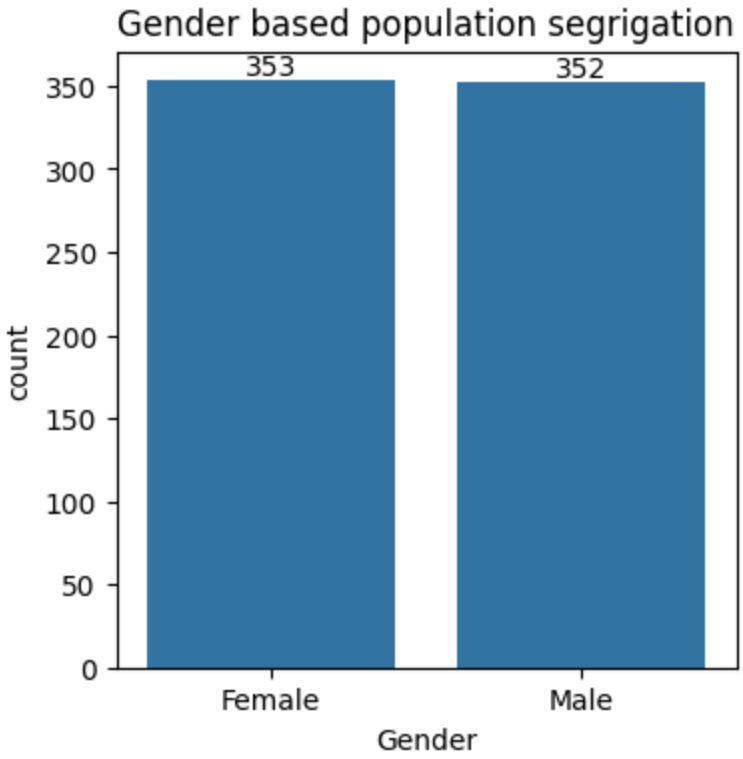
Out[5]:
```
Student_ID                      0
Age                             0
Gender                          0
Academic_Level                  0
Country                         0
Avg_Daily_Usage_Hours           0
Most_Used_Platform              0
Affects_Academic_Performance    0
Sleep_Hours_Per_Night           0
Mental_Health_Score             0
Relationship_Status             0
Conflicts_Over_Social_Media     0
Addicted_Score                  0
dtype: int64
```

In [6]:
```
df.duplicated()
```

Out[6]:
```
0      False
1      False
2      False
3      False
4      False
       ...
700    False
701    False
702    False
703    False
704    False
Length: 705, dtype: bool
```

In [7]:
```
plt.figure(figsize = (4,4))
ax = sns.countplot(x="Gender", data=df)
ax.bar_label(ax.containers[0])
```

```
plt.title("Gender based population segrigation")
plt.show()
```



Total sample size = 705 % of Male population = (352/700)*100 = 49.93 % % of Female population = (353/700)*100 = 50.07 %
This shows our data is balance

In [8]:
```
df1 = df["Country"].value_counts()
df1
```

Out[8]:
```
Country
India          53
USA            40
Canada         34
France         27
Mexico         27
               ..
Oman            1
Afghanistan     1
Syria           1
Yemen           1
Bhutan          1
Name: count, Length: 110, dtype: int64
```

Total Countries in the dataset = 110 Top 5 countries: India USA Canda France Mexico

Q1. What is the average daily social media usage (in hours) across all students?

In [9]:
```
Total_hours = df["Avg_Daily_Usage_Hours"].sum()
Total_population = 705
average_usage = Total_hours/Total_population
print(average_usage)
```

```
4.918723404255319
```

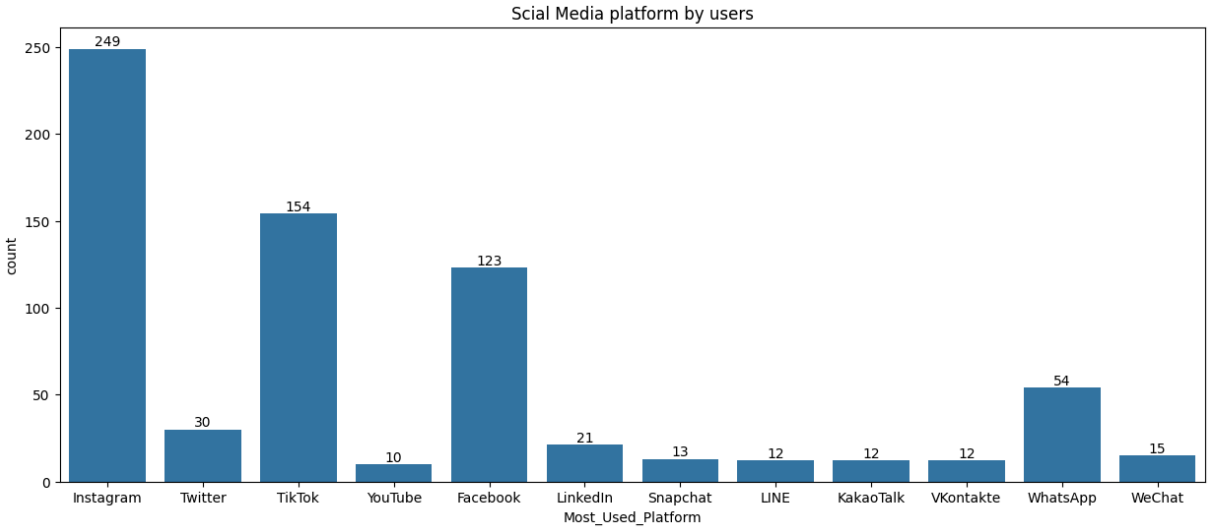
average daily social media usage (in hours) across all students = 4.9 hours

Q2. Which social media platform is the most frequently reported as "Most_Used_Platform"?

In [10]:
```
Most_used_platform = df["Most_Used_Platform"].value_counts()
Most_used_platform
```

Out[10]:
```
Most_Used_Platform
Instagram    249
TikTok       154
Facebook     123
WhatsApp      54
Twitter       30
LinkedIn      21
WeChat        15
Snapchat      13
VKontakte     12
LINE          12
KakaoTalk     12
YouTube       10
Name: count, dtype: int64
```

In [11]:
```
plt.figure(figsize = (15,6))
ax = sns.countplot(x="Most_Used_Platform", data=df)
ax.bar_label(ax.containers[0])
plt.title("Scial Media platform by users")
plt.show()
```



Most Used plateform = Instagram top 5 plateform: Instagrma TikTok Facebook WhatsApp Twitter(X)

Q3. How many students report that social media affects their academic performance (Yes vs. No)?

In [12]:
```
Performance = df["Affects_Academic_Performance"].value_counts()
Performance
```

Out[12]:
```
Affects_Academic_Performance
Yes    453
No     252
Name: count, dtype: int64
```

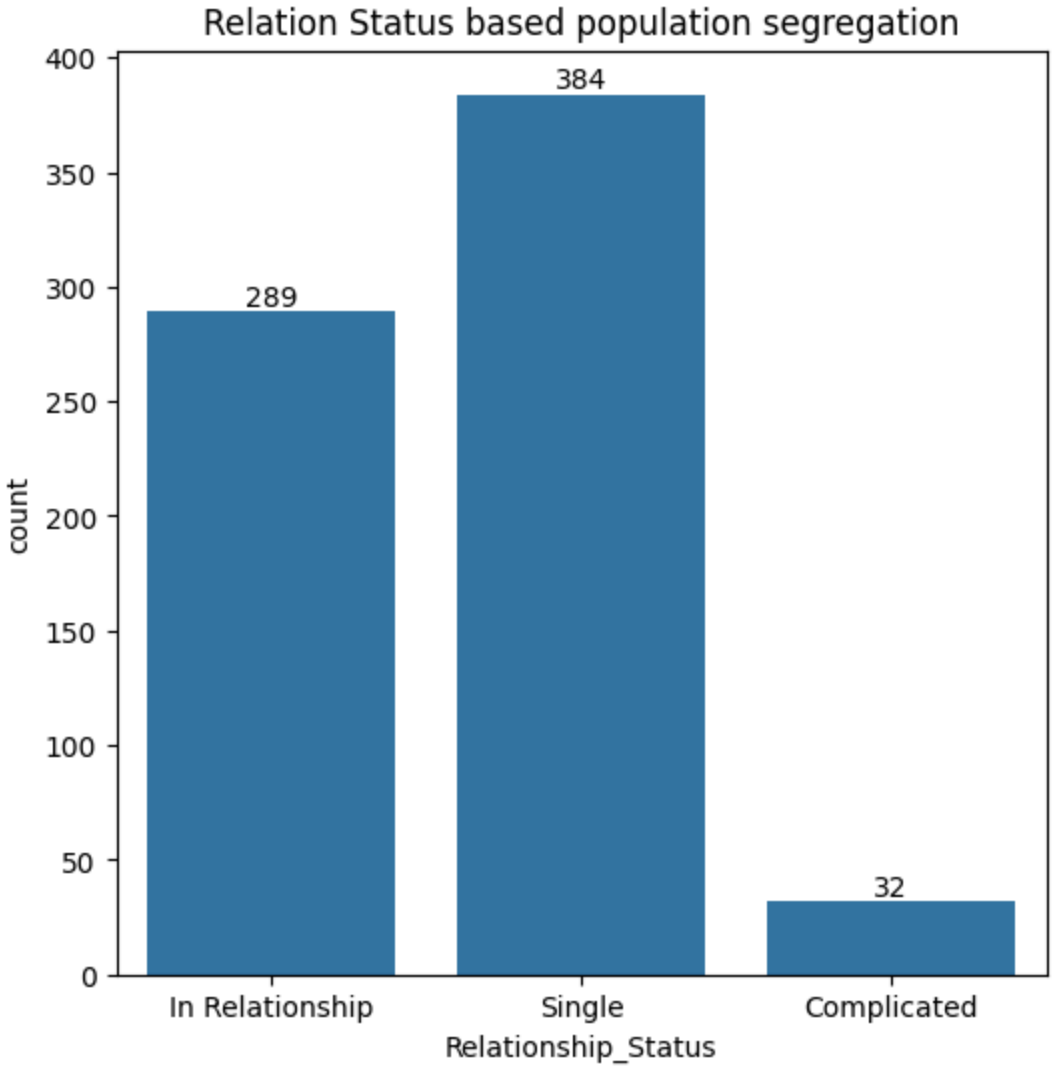
453 students report that their addiction of social media affects their academic performance Their percentage = (453/705)*100 = 64.26%

Q4. What is the distribution of students by Relationship_Status (Single, In Relationship, Complicated)?

In [13]:
```
Relationship = df["Relationship_Status"].value_counts()
Relationship
```

Out[13]:
```
Relationship_Status
Single              384
In Relationship     289
Complicated          32
Name: count, dtype: int64
```

In [14]:
```
plt.figure(figsize = (6,6))
ax = sns.countplot(x="Relationship_Status", data=df)
ax.bar_label(ax.containers[0])
plt.title("Relation Status based population segregation")
plt.show()
```



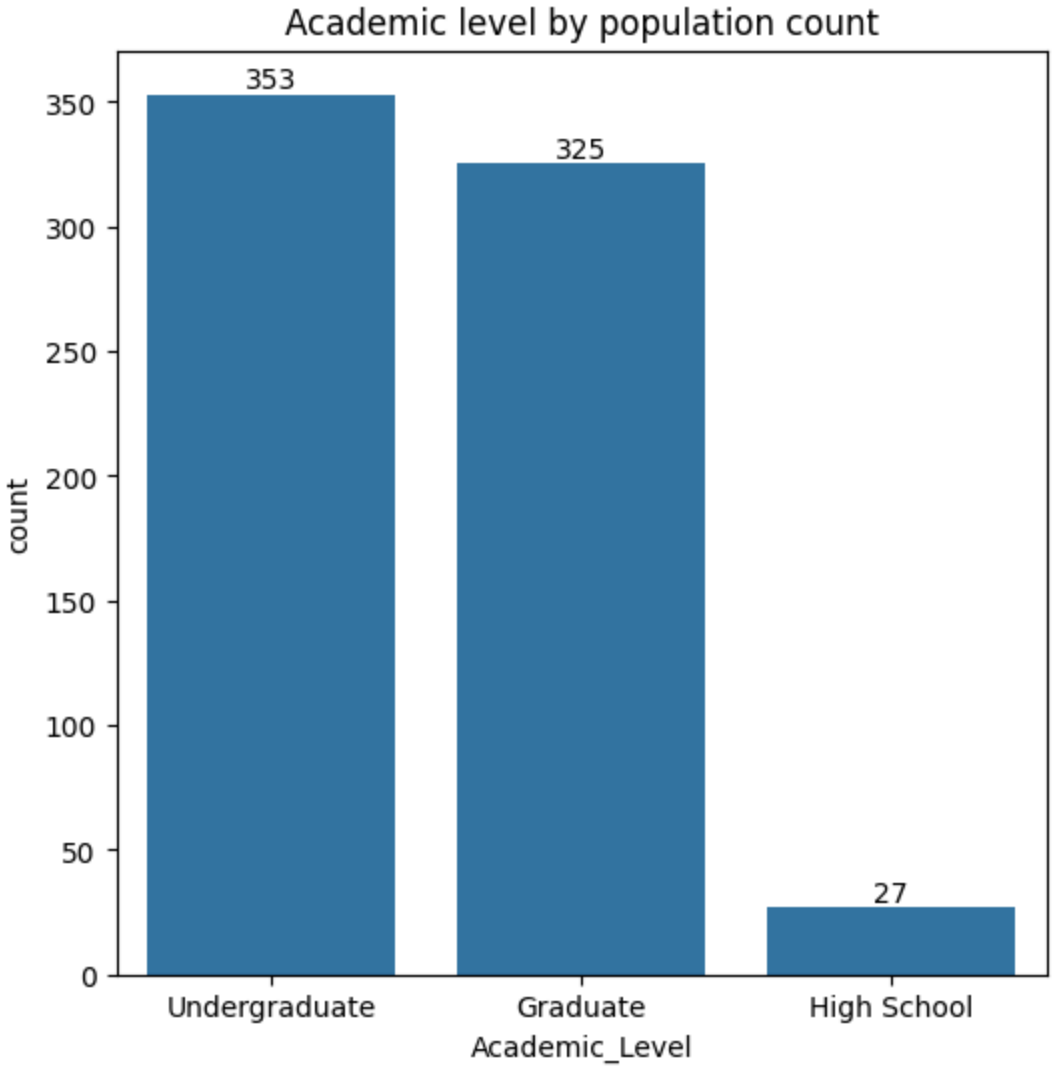
Most of the student in this dataset are single % of singles = (384/705)*100 = 54.47 % of students that are in relationship = (289/705)*100 = 40.99 % of students otherwise (complicated) = 4.54

Q5. What is the distribution of students by Academic_Level ?

In [15]:
```
academicLevel = df["Academic_Level"].value_counts()
academicLevel
```

Out[15]:
```
Academic_Level
Undergraduate    353
Graduate         325
High School       27
Name: count, dtype: int64
```

In [16]:
```
plt.figure(figsize = (6,6))
ax = sns.countplot(x="Academic_Level", data=df)
ax.bar_label(ax.containers[0])
plt.title("Academic level by population count")
plt.show()
```



% of Undergraduates = (353/705)*100 = 50.07 % of Graduates = (325/705)*100 = 46.10 % of High School = (27/705)*100 = 3.83

Q6. Is there a correlation between daily social media usage and sleep duration ?

In [17]:

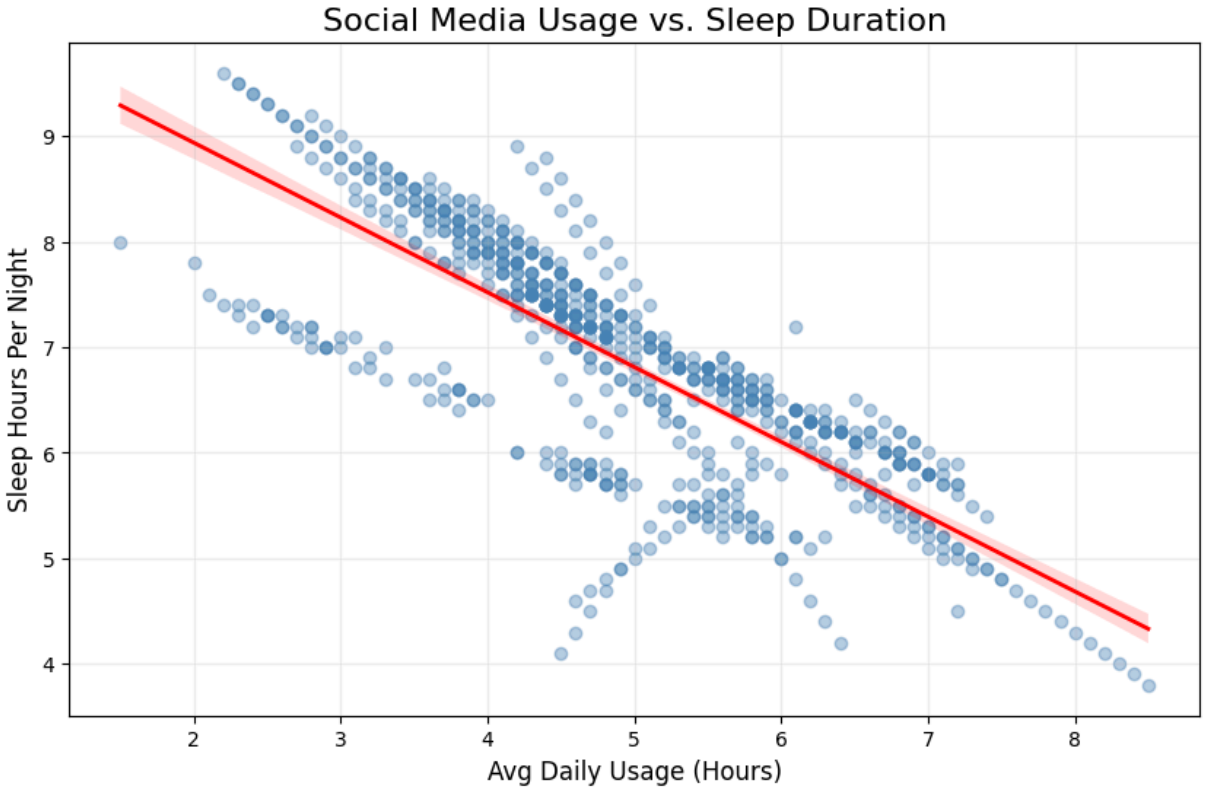
```
from scipy import stats

# Calculate Pearson correlation
corr, p_value = stats.pearsonr(df['Avg_Daily_Usage_Hours'], df['Sleep_Hours_
print(f"Pearson r: {corr:.3f}, p-value: {p_value:.4f}")

# Create scatter plot
plt.figure(figsize=(10, 6))
sns.regplot(
    x='Avg_Daily_Usage_Hours',
    y='Sleep_Hours_Per_Night',
    data=df,
    scatter_kws={'alpha': 0.4, 'color': 'steelblue'},
    line_kws={'color': 'red', 'linewidth': 2}
)
```

```
# Labels and title
plt.title('Social Media Usage vs. Sleep Duration', fontsize=16)
plt.xlabel('Avg Daily Usage (Hours)', fontsize=12)
plt.ylabel('Sleep Hours Per Night', fontsize=12)
plt.grid(alpha=0.2)
plt.show()
```

```
Pearson r: -0.791, p-value: 0.0000
```



In [18]:
```
from scipy import stats

# Calculate linear regression parameters
slope, intercept, r_value, p_value, std_err = stats.linregress(
    df['Avg_Daily_Usage_Hours'],
    df['Sleep_Hours_Per_Night']
)
print(slope,intercept)
```

```
-0.7085018080615694 10.353860595482416
```

Strong negative correlation (r = -0.791) High social media users (6+ hrs/day) sleep 1-3 hours less than low users "Blue light" effect: Screen exposure reduces melatonin, delaying sleep Behavioral displacement: Time spent on apps replaces sleep time Real-world insight: Reducing social media use by 1 hour may increase sleep by ~40 minutes (slope=0.7)

Q7. Which country has the highest average Addicted_Score? (Limit to top 5 countries)

In [21]:
```
Countries = df["Country"].value_counts()
Countries
```

Out[21]:
```
Country
India          53
USA            40
Canada         34
France         27
Mexico         27
               ..
Oman            1
Afghanistan     1
Syria           1
Yemen           1
Bhutan          1
Name: count, Length: 110, dtype: int64
```

In [23]:
```
valid_countries = Countries[Countries >= 10].index
filtered_df = df[df['Country'].isin(valid_countries)]
filtered_df
```

Out[23]:

	Student_ID	Age	Gender	Academic_Level	Country	Avg_Daily_Usage_I
0	1	19	Female	Undergraduate	Bangladesh	
1	2	22	Male	Graduate	India	
2	3	20	Female	Undergraduate	USA	
3	4	18	Male	High School	UK	
4	5	21	Male	Graduate	Canada	
...	...	...	...	...	...	
700	701	20	Female	Undergraduate	Italy	
701	702	23	Male	Graduate	Russia	
702	703	21	Female	Undergraduate	China	
703	704	24	Male	Graduate	Japan	
704	705	19	Female	Undergraduate	Poland	

570 rows × 13 columns

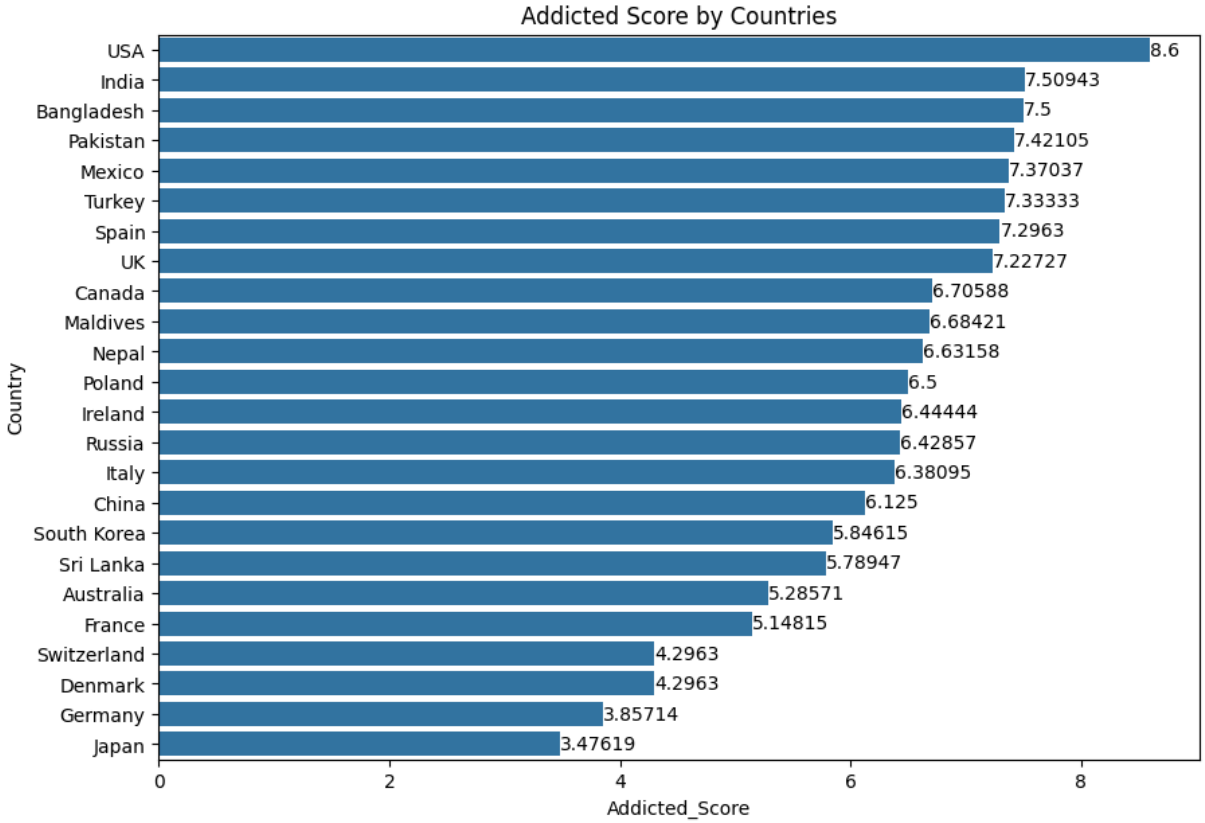
In [26]:
```
country_scores = filtered_df.groupby('Country')['Addicted_Score'].mean().res
top_countries = country_scores.sort_values('Addicted_Score', ascending=False
top_countries
```

Out[26]:

	Country	Addicted_Score
23	USA	8.600000
7	India	7.509434
1	Bangladesh	7.500000
14	Pakistan	7.421053
12	Mexico	7.370370
21	Turkey	7.333333
18	Spain	7.296296
22	UK	7.227273
2	Canada	6.705882
11	Maldives	6.684211
13	Nepal	6.631579
15	Poland	6.500000
8	Ireland	6.444444
16	Russia	6.428571
9	Italy	6.380952
3	China	6.125000
17	South Korea	5.846154
19	Sri Lanka	5.789474
0	Australia	5.285714
5	France	5.148148
20	Switzerland	4.296296
4	Denmark	4.296296
6	Germany	3.857143
10	Japan	3.476190

In [31]:

```
plt.figure(figsize = (10,7))
ax = sns.barplot(x="Addicted_Score", y="Country", data=top_countries)
ax.bar_label(ax.containers[0])
plt.title("Addicted Score by Countries")
plt.show()
```



USA have the highest Addictive Score followed by India and Bangladesh

Q8. Do students who report academic performance issues (Affects_Academic_Performance = "Yes") have higher average Addicted_Scores than those who don't?

In [32]:
```
Performance = df["Affects_Academic_Performance"].value_counts()
Performance
```

Out[32]:
```
Affects_Academic_Performance
Yes    453
No     252
Name: count, dtype: int64
```

In [34]:
```
AAS= df.groupby('Affects_Academic_Performance')['Addicted_Score'].mean().res
AAS
```

Out[34]:

	Affects_Academic_Performance	Addicted_Score
0	No	4.595238
1	Yes	7.461369

yes, students whome academic performance are affected have higher addicted score (7.461369)

Q9. How does mental health (Mental_Health_Score) differ between genders?

In [35]:
```
Mental_Health = df.groupby('Gender')['Mental_Health_Score'].mean().reset_inc
Mental_Health
```

Out[35]:

	Gender	Mental_Health_Score
0	Female	6.175637
1	Male	6.278409

Overall both genders has similar Mental health score

Q10. Do students in "Complicated" relationships experience more conflicts over social media (Conflicts_Over_Social_Media) than those "In Relationship" or "Single"?

In [39]:
```
Conflicts = df.groupby('Relationship_Status')['Conflicts_Over_Social_Media']
Conflicts
```

Out[39]:

	Relationship_Status	Conflicts_Over_Social_Media
0	Complicated	3.031250
1	In Relationship	2.761246
2	Single	2.901042

yes, average conflicts over social media is more in case of students who have complicated relationship