Universidad Autónoma de Baja California

Programación para Ciencia de Datos
Profesor: Raul Casillas Figueroa

Proyecto Final:
Salud mental y redes sociales

368413, Kevy Leonardo Oviedo Guia

Fecha: 06 / Junio / 2024

Proyecto Final:
Salud mental y redes sociales

El presente caso trata sobre la relación que existe entre la salud mental de las personas y su uso de las redes sociales. Debido a que las redes sociales se han vuelto de los medio de comunicación más utilizados en las últimas décadas y la facilidad que tienen en su uso, entonces es importante tomar en cuenta si tienen efectos negativos en nuestros comportamientos y salud mental. Y, saber si hay medidas preventivas ante estos problemas que se puedan ocasionar por su uso.

Para analizar el caso, se requieren datos generales, de uso de redes sociales y de conductas de personas que utilicen redes sociales. En este caso se usarán datos como: edad, género, tiempo en redes sociales y el cómo se sienten las personas en cuanto a concentración, estado del ánimo, problemas de salud, autoestima, etc. Para así poder obtener conclusiones sobre si hay alguna relación entre uso de redes sociales y salud mental.

Para obtener los datos, se hizo uso de un dataset titulado “Social Media and Mental Health” de la página *Kaggle*. Este dataset contiene información relevante sobre el caso que se está analizando. Algunas columnas que tiene son: edad, género, tiempo en redes sociales, sentimientos de depresión, problemas de sueño, comparación con otros, búsqueda de validación, problemas de concentración, entre otros. En el caso de las columnas de cómo se sienten y problemas que tengan, se usa una escala del 1 al 5 para determinarlo, siendo que 1 es que se sienten bien o no tienen problemas con eso y 5 que tienen problemas con ello.

Para realizar el código se hizo un notebook utilizando jupyter.

Primero, se importaron las librerías pandas, numpy, matplotlib y seaborn. Se genero el data frame y se eliminaron algunos datos del data frame, como columnas que no se van a utilizar y personas a los que no apliquen las preguntas

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
```

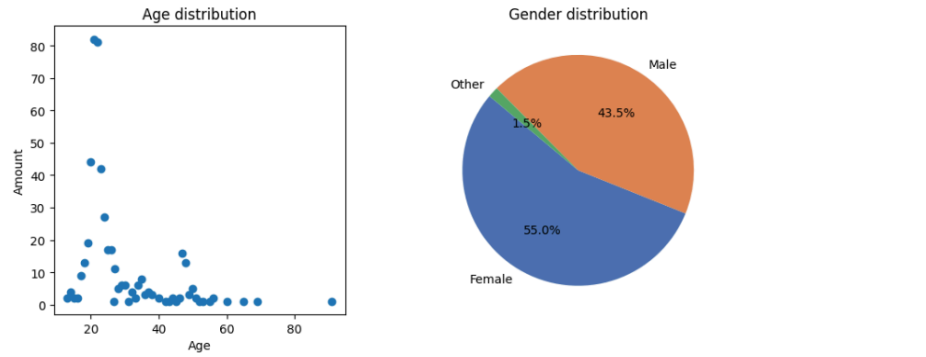
```
# Generar dataFrame
SMhealth_df = pd.read_csv('smmh.csv')

# Eliminar columnas que no se usaran
SMhealth_df = SMhealth_df.drop(['Timestamp', '5. What type of organizations are you affiliated with?', ], axis=1)
```

```
#Eliminar a todos los que no usen redes sociales
SMhealth_df = SMhealth_df[SMhealth_df['6. Do you use social media?'] != 'No']
# Renombrar 'Less than an hour' a '.Less than an hour', para que salga en orden correcto
SMhealth_df['8. What is the average time you spend on social media every day?'] = SMhealth_df['8. What is the average time you sp
```

Para la primera serie de gráficas, se implementó el siguiente código. El cual busca obtener datos generales de las personas del dataset,como la distribución de edades, el porcentaje de hombres y mujeres, sus ocupaciones, y las redes sociales que más se están utilizando. Se hizo uso de los subplots para poner las 4 gráficas y se usaron gráficas de barras, pie y scatter.

```
# Distribucion de edades que usan redes sociales
AgeTime_df = SMhealth_df[['1. What is your age?', '8. What is the average time you spend on social media every day?']]
age_counts = AgeTime_df['1. What is your age?'].value_counts().sort_index()

# Distribucion de generos del dataset
not_male_female = ~SMhealth_df['2. Gender'].isin(['Male', 'Female'])
SMhealth_df.loc[not_male_female, '2. Gender'] = 'Other'
gender_count = SMhealth_df['2. Gender'].value_counts().sort_index()

# Distribcion deocupacion del dataset
occupation_count = SMhealth_df['4. Occupation Status'].value_counts().sort_index()

# Usuarios de red social en dataset
socialM_df = SMhealth_df['7. What social media platforms do you commonly use?'].str.split(',').explode().str.strip()
socialM_counts = SocialM_df.value_counts().reset_index()
socialM_counts.columns = ['Social Media', 'Users']

# hacer subpltos de 2x2
fig, axes = plt.subplots(2, 2, figsize=(10,10))

#Hacer plot (0,0) de distribucion de edades con scatter plot
axes[0,0].set_xlabel('Age')
axes[0,0].set_ylabel('Amount')
axes[0,0].set_title('Age distribution')
axes[0,0].scatter(age_counts.index, age_counts.values)

# Hacer plot (0,1) de distribucion de generos con pie
axes[0,1].pie(gender_count.values, labels=gender_count.index, autopct='%1.1f%%', startangle=140, colors=sns.color_palette('deep'))
axes[0,1].set_title('Gender distribution')

# Hacer plot (1,0) de usuarios en cada red social con barplot
sns.barplot(x='Social Media', y='Users', data=socialM_counts,hue='Social Media',legend=True, palette='pastel', ax=axes[1,0])
axes[1,0].set_xticks([])
axes[1,0].set_title('Users in each Social Media')
axes[1,0].legend(title='Social Media', loc='upper right')

# Hacer plot (1,1) de distribucion de ocupacion con pie
axes[1,1].pie(occupation_count.values, labels=occupation_count.index, autopct='%1.1f%%', startangle=140, colors=sns.color_palette('pastel'))
axes[1,1].set_title('Occupation distribution')
```

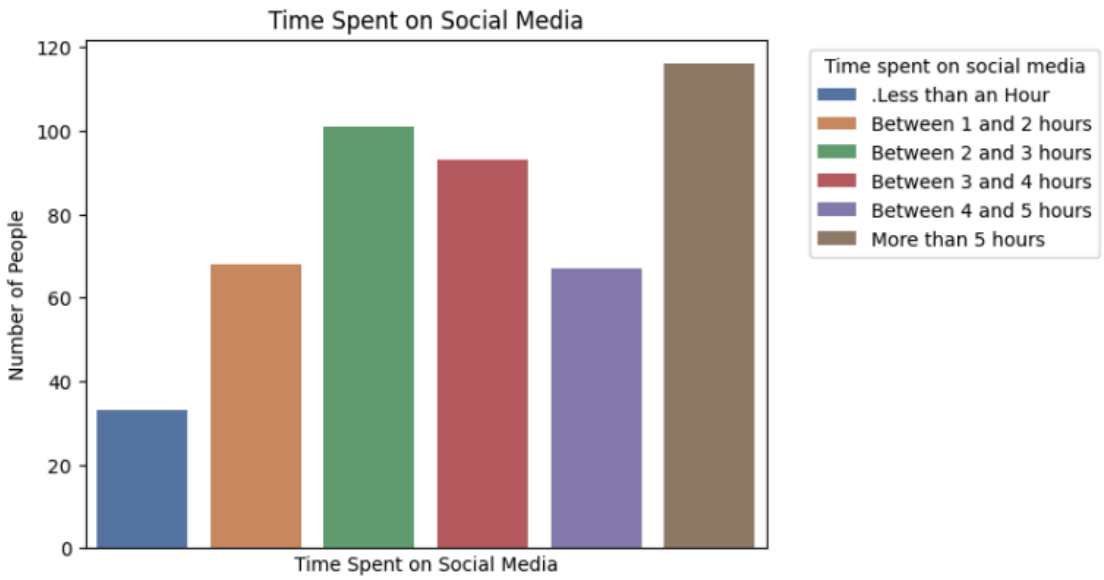
Las gráficas obtenidas fueron:
Se utilizaron pie porque son buenas para representar porcentajes. Scatter porque muestra de forma visible la edad y la cantidad de personas con esa edad. De barras para diferenciar entre las categorías y su cantidad de usuarios



El pedazo de código mostrado, busca contar el número de personas que usan redes sociales durante diversos lapsos de tiempo, por ejemplo ‘menos de una hora’, ‘entre 1 y 2 horas’, etc. Se grafica utilizando barplot

```
# crear dataframe timeSpent donde se tiene el tiempo que se usa una red social y cuatnas personas la usan durante ese tiempo
time_counts = AgeTime_df['8. What is the average time you spend on social media every day?'].value_counts().sort_index()
timeSpent_df = time_counts.reset_index()
timeSpent_df.columns = ['Time spent on social media','Number of people']

# hacer plot del timeSpent_df con barplot
sns.barplot(x='Time spent on social media', y='Number of people', data=timeSpent_df ,legend=True, palette='deep', hue = 'Time spent on social media')
plt.xticks([])

plt.xlabel('Time Spent on Social Media')
plt.ylabel('Number of People')
plt.title("Time Spent on Social Media")

plt.legend(title='Time spent on social media', bbox_to_anchor=(1.05, 1), loc='upper left')
```

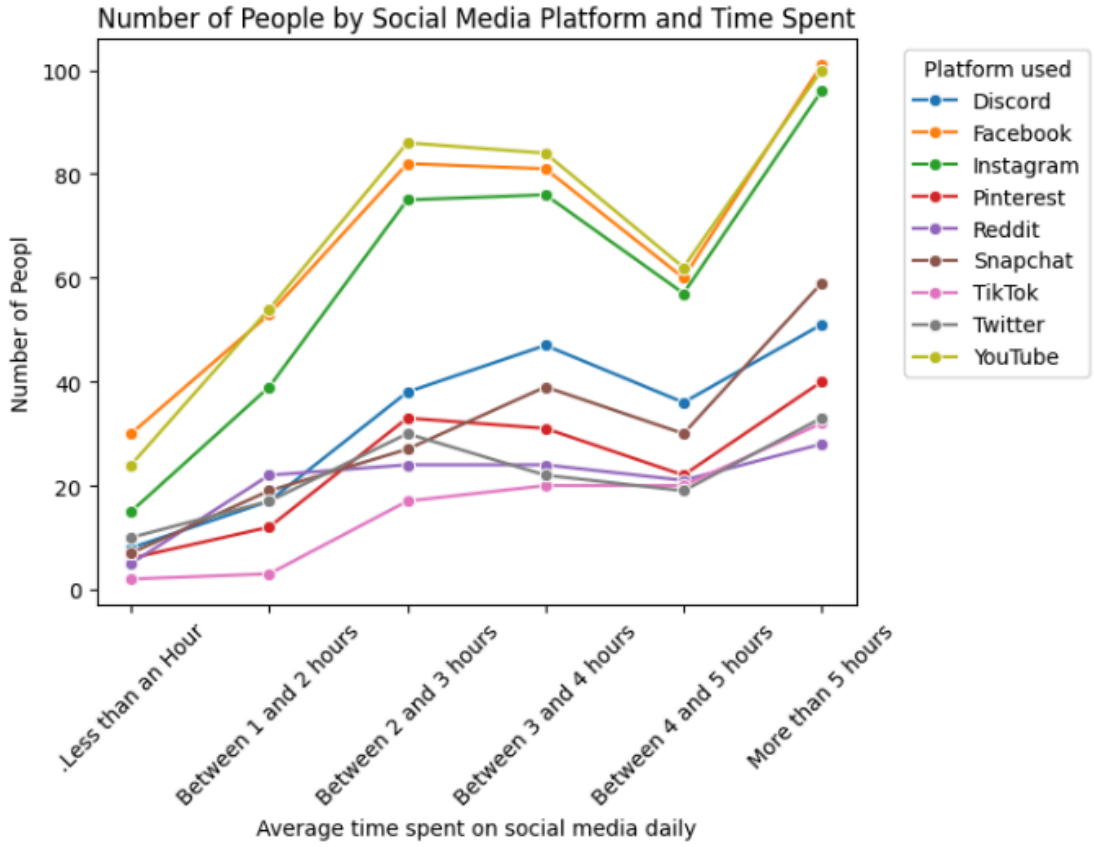
La gráfica obtenida es una gráfica de barras con una leyenda usando colores para representar cada rango de tiempo. Se utilizaron barras porque se cuenta el número de personas por categoría, entonces es fácil diferenciar cada aspecto.



Se hizo un dataframe llamado usageMedia_df en el cual se separaron los valores separados por comas de ‘Platforms used’ y se les dio un renglón a cada uno. Después, se hizo un count para contar el número de personas en una plataforma e intervalo de tiempo. Se hizo una gráfica tipo lineplot para representar los datos

```
# Generar un dataframe al separar las redes sociales de la columna con las plataformas utilziadas e indicar el rango de horas que son utilizadas.
usageMedia_df = SMhealth_df[['8. What is the average time you spend on social media every day?','7. What social media platforms do you commonly use?']]
usageMedia_df = usageMedia_df.rename(columns={'8. What is the average time you spend on social media every day?':'Average time spent on social media dai
media_df = usageMedia_df['Platform used'].str.split(', ', expand=True).stack().reset_index(level=1, drop=True).rename('Platform used')
usageMedia_df = usageMedia_df.drop('Platform used', axis=1).join(media_df)
usageMedia_df = usageMedia_df.reset_index(drop=True)

usageMedia_df = usageMedia_df.sort_values(by='Average time spent on social media daily')
usageMedia_df = usageMedia_df.groupby(['Average time spent on social media daily', 'Platform used']).size().reset_index(name='count')

sns.lineplot(x='Average time spent on social media daily', hue='Platform used',y='count', data=usageMedia_df, marker='o')

plt.xlabel("Average time spent on social media daily")
plt.ylabel("Number of Peopl")
plt.title("Number of People by Social Media Platform and Time Spent")

plt.xticks(rotation=45)
plt.legend(title='Platform used', bbox_to_anchor=(1.05, 1), loc='upper left')
#plt.tight_layout()
```

La grafica obtenida fue del tipo lineplot. El eje x muestra el tiempo que se usa en redes sociales, el hue son las diversas plataformas de redes sociales que existen y el eje y es la cantidad de personas que usan una red social en un intervalo de tiempo.



Se realizaron 3 tablas tipo catplot pero con diferentes datos. En estas se utilizaron datos de cuánto usan redes sociales, y el valor que le dieron a sus problemas de distracción, calidad de sueño, y depresión. Entonces, el eje x indica su respuesta (1-5) y se agrupan las tablas del tiempo de uso, y el eje y es la cantidad de personas que cumplen esas condiciones

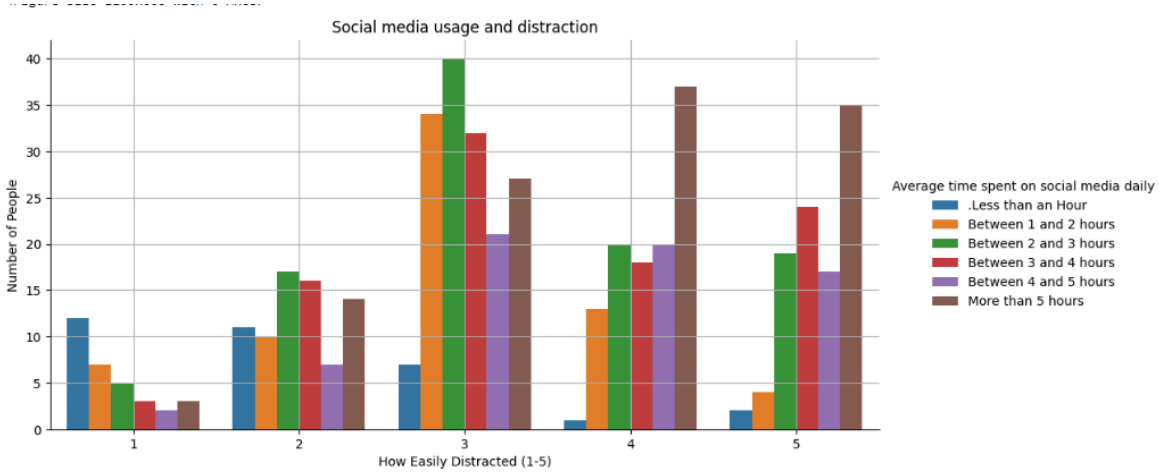
```
# Datos para comparar el tiempo de uso de una red social y la distraccion de las personas.
SMhealth_df['8. What is the average time you spend on social media every day?'] = SMhealth_df['8. What is the average time you spend on social media eve
distractionGrouped_df = SMhealth_df.groupby(['12. On a scale of 1 to 5, how easily distracted are you?', '8. What is the average time you spend on socia
plt.figure(figsize=(12,6))
distractionGrouped_df=distractionGrouped_df.rename(columns={'8. What is the average time you spend on social media every day?':'Average time spent on so

#Hacer plot del dataframe con catplot
sns.catplot(x='12. On a scale of 1 to 5, how easily distracted are you?', y='count', hue='Average time spent on social media daily', data=distractionGro
plt.xlabel('How Easily Distracted (1-5)')
plt.ylabel('Number of People')
plt.title('Social media usage and distraction')
plt.xticks(ticks=[0, 1, 2, 3, 4], labels=['1', '2', '3', '4', '5'])
plt.grid(True)
```

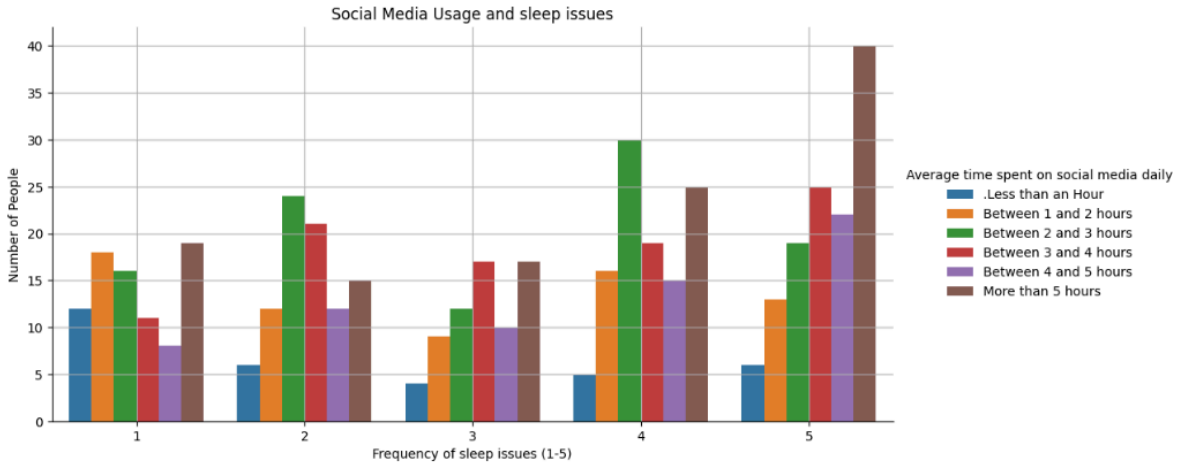
```python
# Datos para comparar el tiempo de uso de una red social y problemas del sueno de las personas.
sleepGrouped_df = SMhealth_df.groupby(['20. On a scale of 1 to 5, how often do you face issues regarding sleep?', '8. What is the average time you spend
plt.figure(figsize=(12,6))
sleepGrouped_df=sleepGrouped_df.rename(columns={'8. What is the average time you spend on social media every day?':'Average time spent on social media d

# Hacer plot del dataframe con catplot
sns.catplot(x='20. On a scale of 1 to 5, how often do you face issues regarding sleep?', y='count', hue='Average time spent on social media daily', data
plt.xlabel('Frequency of sleep issues (1-5)')
plt.ylabel('Number of People')
plt.title('Social Media Usage and sleep issues')
plt.xticks(ticks=[0, 1, 2, 3, 4], labels=['1', '2', '3', '4', '5'])
plt.grid(True)
```

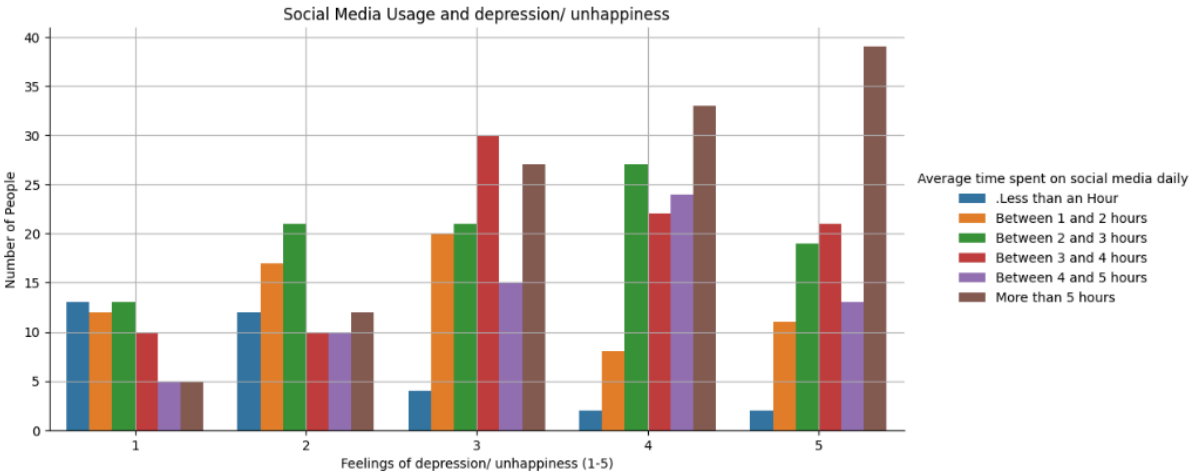
```python
# Datos para comparar el tiempo de uso de una red social y depresion
depressionGrouped_df = SMhealth_df.groupby(['18. How often do you feel depressed or down?', '8. What is the average time you spend on social media every
plt.figure(figsize=(12,6))
depressionGrouped_df=depressionGrouped_df.rename(columns={'8. What is the average time you spend on social media every day?':'Average time spent on soci

#Hacer plot del dataframe con catplot
sns.catplot(x='18. How often do you feel depressed or down?', y='count', hue='Average time spent on social media daily', data=depressionGrouped_df, kind
plt.xlabel('Feelings of depression/ unhappiness (1-5)')
plt.ylabel('Number of People')
plt.title('Social Media Usage and depression/ unhappiness')
plt.xticks(ticks=[0, 1, 2, 3, 4], labels=['1', '2', '3', '4', '5'])
plt.grid(True)
```

Las tablas son catplot porque se busca mostrar tres diferentes datos. La primera gráfica es para medir el uso de redes sociales con distracción, al mostrar el número de gente con el tiempo que usa redes sociales y que tan distraídos son. La segunda y tercera miden el uso de redes con problemas para dormir y de depresión, y también muestra el número de gente como en la primera gráfica

.



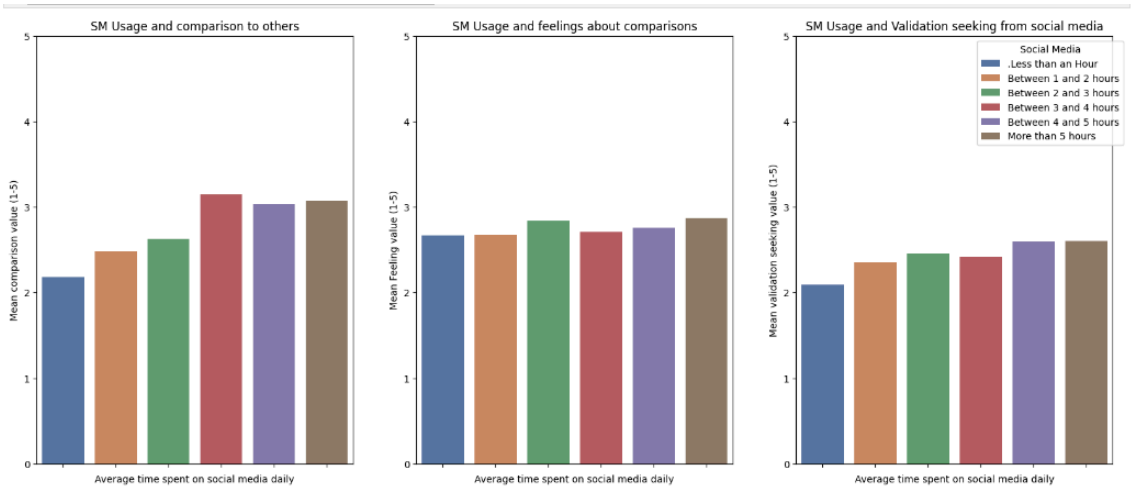




Para estas gráficas, se hizo un nuevo dataset comparison_df, donde solo están las columnas de tiempo de uso de redes sociales, frecuencia de comparación con otros, sentimientos de comparación, búsqueda de validez.Se hizo un groupby y se obtuvieron 3 tablas para cada comparación. En este caso, en lugar de obtener el número de personas en la categoría, se realizó un promedio de los valores que se indicaron para cada categoría. Las gráficas se hicieron con el subplot y las 3 fueron de barras con la misma leyenda de colores.

```python
comparison_df = SMhealth_df.rename(columns={'15. On a scale of 1-5, how often do you compare yourself to other successful people through the use of soci
comparison_df = comparison_df.rename(columns={'8. What is the average time you spend on social media every day?':'Average time spent on social media dai

comparison_Others = comparison_df.groupby('Average time spent on social media daily')['Comparison to others'].mean()
comparison_Feelings = comparison_df.groupby('Average time spent on social media daily')['Feelings about comparisons'].mean()
comparison_Validation = comparison_df.groupby('Average time spent on social media daily')['Validation seeking from social media'].mean()

#Crear subplots de 1 renglon y 3 columnas
fig, axes = plt.subplots(1, 3, figsize=(20, 8))

# hacer barplot de uso de redes sociales y cuanto se comapran con otros
sns.barplot(x=comparison_Others.index, y=comparison_Others.values, ax=axes[0],legend=False, palette='deep', hue=comparison_Validation.index)
axes[0].set_xlabel('Average time spent on social media daily')
axes[0].set_ylabel('Mean comparison value (1-5)')
axes[0].set_title('SM Usage and comparison to others')
axes[0].set_ylim(0,5)

# Hacer barplot  de uso de redes sociales y el como son los sentimientos cuando se comapran
sns.barplot(x=comparison_Feelings.index, y=comparison_Feelings.values, ax=axes[1],legend=False, palette='deep', hue=comparison_Validation.index)
axes[1].set_xlabel('Average time spent on social media daily')
axes[1].set_ylabel('Mean Feeling value (1-5)')
axes[1].set_title('SM Usage and feelings about comparisons')
axes[1].set_ylim(0,5)

# Hacer barplot de uso de redes sociales y la validacion que buscan de redes sociales
sns.barplot(x=comparison_Validation.index, y=comparison_Validation.values, ax=axes[2],legend=True, palette='deep', hue=comparison_Validation.index)
axes[2].set_xlabel('Average time spent on social media daily')
axes[2].set_ylabel('Mean validation seeking value (1-5)')
axes[2].set_title('SM Usage and Validation seeking from social media')
axes[2].set_ylim(0,5)
axes[2].legend(title='Social Media',bbox_to_anchor=(1.05,1), loc='upper right')

# Remover los valroes de los indices del eje x
for i in range(0,3):
    axes[i].set_xticks([0,1,2,3,4,5])
    axes[i].set_xticklabels([])
```

Las 3 tablas se encuentran dentro de un mismo plot y son de barras. Se decidió esto porque todas dan información sobre comparación que realizan las personas con otros y relacionado al tiempo que pasan en redes sociales. Tienen leyendas para identificar el rango de tiempo con diferentes colores en las tablas.



Como conclusión, se observa que hay una correlación muy marcada entre el tiempo de uso de las redes sociales y problemas en la salud mental de los individuos. Mientras que los que pasaban más de 5 horas al día usando redes contestaban que experimentaban diversos problemas, los que menos las utilizaban marcaban con mayor regularidad que no se veían afectados. De igual forma, se obtuvo que hay una tendencia por compararse con la gente en las redes sociales y que esta es normalmente negativa en todo rango de tiempo de uso.
Otro punto importante por mencionar es que los datos están un poco sesgados en cuanto al grupo de edad, ya que la mayoría de gente del dataset eran personas de 18 a 25 años de edad, por lo que los comportamientos pueden ser únicamente para este rango de edad y no se pueden generalizar para gente menor o mayor.