ANÁLISIS EXPLORATORIO DEL DATASET TITANIC

Análisis exploratorio del Dataset Titanic¶

Introducción¶

El naufragio del Titanic es uno de los desastres marítimos más conocidos de la historia, y el dataset asociado proporciona una oportunidad única para analizar los factores que pudieron influir en la supervivencia de los pasajeros. Este proyecto tiene como objetivo explorar y visualizar los datos del Titanic para identificar patrones y tendencias que expliquen la probabilidad de supervivencia.

Preparación de Datos¶

Limpieza de datos¶

Durante la preparación del dataset:

- Se reemplazaron los valores nulos en la columna Age con la media de todas las edades.
- Se modificaron las variables categóricas como Sex y Embarked para convertirlas en valores numéricos.
- Se agregó una nueva columna llamada Tamaño_Familia, que combina las columnas SibSp (hermanos/esposos a bordo) y Parch (padres/hijos a bordo).

Selección de variables¶

Las variables elegidas para el analisis son:

- Pclass (Clase del pasajero: 1^{a},2^{a},3^{a}).
- Sex (Género del pasajero).
- Age (Edad del pasajero).
- SibSp y Parch (Número de familiares a bordo).
- Fare (Tarifa pagada por el pasajero).
- Embarked (Puero de embarque: C= Cherburgo, Q= Queenstown, S= Southampton).

Análisis Exploratorio¶

Supervivencia por género¶

El análisis reveló que las mujeres tuvieron una mayor probabilidad de supervivencia que los hombres. **Tasa de supervivencia por género:**

- Mujeres: 74.2%
- Hombres: 18.9%

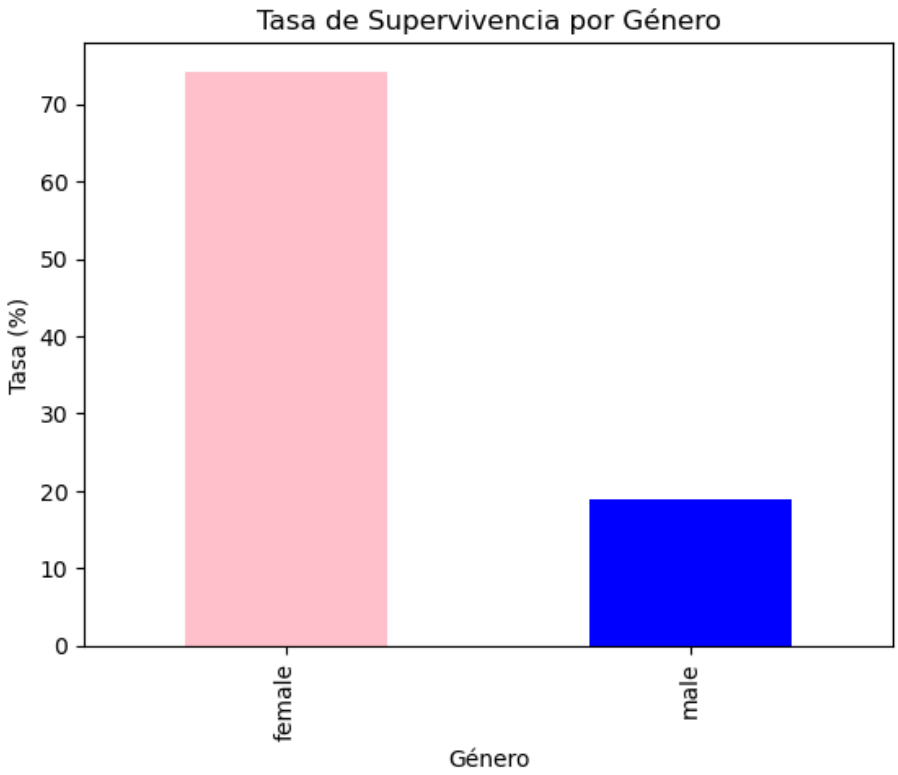
In [16]:

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

ruta= "C:/Users/sohai/Downloads/train.csv"
df =pd.read_csv(ruta)

supervivencia_genero = df.groupby('Sex')['Survived'].mean() * 100

supervivencia_genero.plot(kind='bar', color=['pink', 'blue'])
plt.title("Tasa de Supervivencia por Género")
plt.ylabel("Tasa (%)")
plt.xlabel("Género")
plt.show()
```



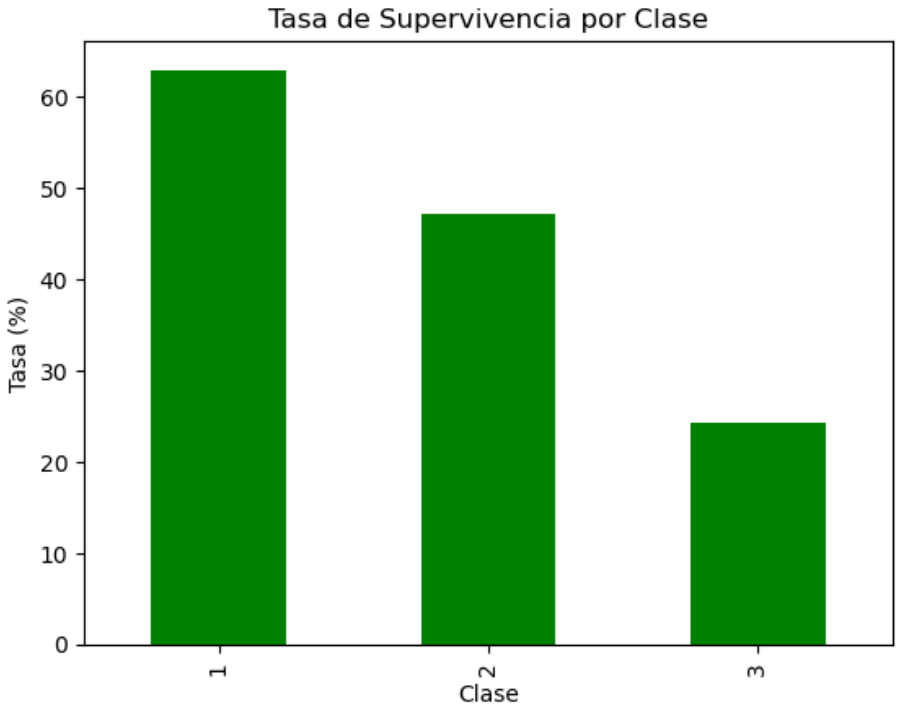
Supervivencia por clase¶

La clase también jugó un papel importante en la supervivencia:

- Primera clase: 62.9%
- Segunda clase: 47.3%
- Tercera clase: 24.2%

In [5]:

```
# Supervivencia por clase
supervivencia_clase = df.groupby('Pclass')['Survived'].mean() * 100
supervivencia_clase.plot(kind='bar', color='green')
plt.title("Tasa de Supervivencia por Clase")
plt.ylabel("Tasa (%)")
plt.xlabel("Clase")
plt.show()
```



Distribución de edades¶

La mayoría de los pasajeros tenía entre 20 y 40 años, y los niños (edad < 16) tuvieron mayores probabilidades de supervivencia.

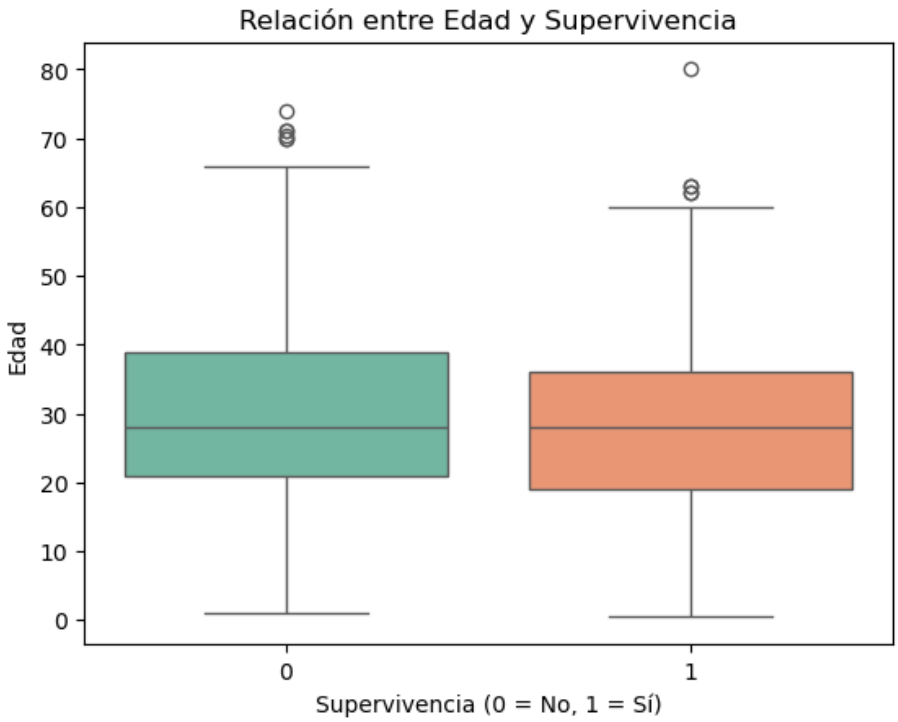
In [11]:

```
# Boxplot de edades por supervivencia
sns.boxplot(data=df, x='Survived', y='Age', palette="Set2")
plt.title("Relación entre Edad y Supervivencia")
plt.xlabel("Supervivencia (0 = No, 1 = Sí)")
plt.ylabel("Edad")
plt.show()
```

```
C:\Users\sohai\AppData\Local\Temp\ipykernel_6448\1466666439.py:2: FutureWarning:

Passing `palette` without assigning `hue` is deprecated and will be removed in v0.14.0. Assi

  sns.boxplot(data=df, x='Survived', y='Age', palette="Set2")
```



Correlaciones¶

Se creó un mapa de calor para analizar la relación entre las variables. Las más correlacionadas con la supervivencia fueron Pclass, Sex y Age.

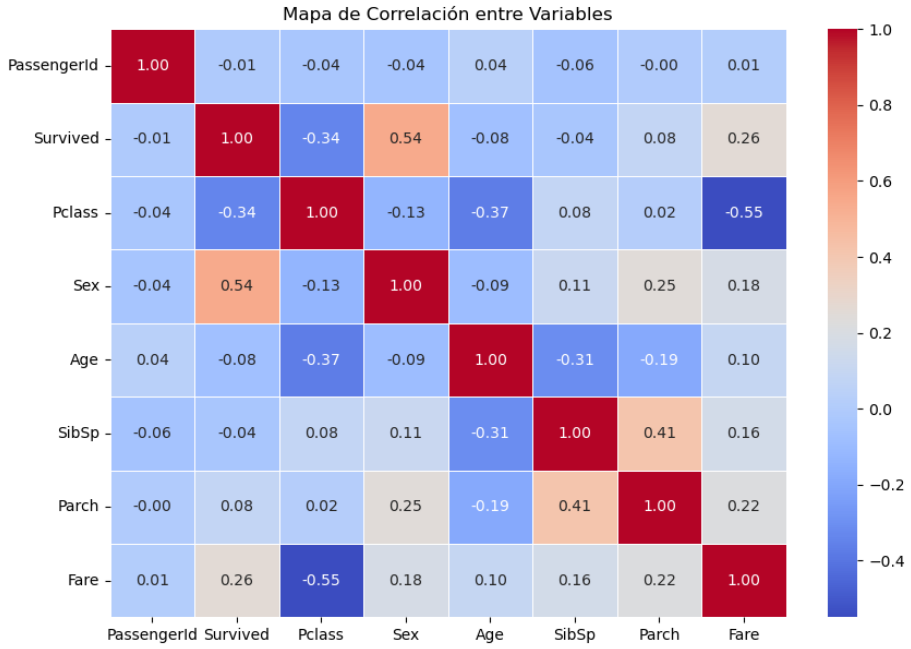
In [14]:

```
df['Sex'] = df['Sex'].map({'male': 0, 'female': 1})
datos_numericos = df.select_dtypes(include=['number'])

correlacion = datos_numericos.corr()

# Visualizar con un mapa de calor
import seaborn as sns

plt.figure(figsize=(10, 7))
sns.heatmap(correlacion, annot=True, cmap="coolwarm", fmt=".2f", linewidths=0.5)
plt.title("Mapa de Correlación entre Variables")
plt.show()
```



Conclusiones¶

A partir del análisis:

1. Las mujeres y los pasajeros de primera clase tuvieron mayores probabilidades de supervivencia.
2. La edad fue un factor significativo: los niños tuvieron más posibilidades de sobrevivir.
3. Los pasajeros que viajaron desde Cherburgo (Embarked = C) también tuvieron una tasa de supervivencia ligeramente mayor.

Próximos pasos¶

Para continuar mejorando este análisis, podrían explorarse las siguientes direcciones:

- Implementar modelos predictivos(como regresión logística o árboles de decisión) para predecir la superviviencia.
- Ampliar el análisis para incluir otras variables derivadas.
- Visualizaciones más detalladas para comunicar insights a audiencias no técnicas.