Оцінювання даної роботи буде ураховувати якість ваших візуалізацій **(**вони можуть незначно відрізнятися від прикладів**,** але повинні змістовно відображати суть задачі**)** та опис отриманих вами результатів**.** При наявності ЛИШЕ візуалізацій без роз'яснень**,** завдання буде оцінена максимум у половину балів**.**

Під час виконання вам дозволено користуватися будь-якими зручними для вас **Python** бібліотеками для візуалізації**.**

I. Receiving Data.

In [2]:

```
# cell for imports. All imports must go here. matplotlib and seaborn
import matplotlib.pyplot as plt
import seaborn as sns
import pandas as pd
```

1. **(2**б**)** Для виконання цієї лабораторної роботи Вам необхідно скористатися вибіркою **titanic (** train.csv**).** Зчитайте його та збережіть його у змінну**,** щоб у подальшому її використати для візуалізацій**.** Виведіть інформацію про кожну колонку**,** її індекс**,** тип та кількість непропущених значень та перші кілька рядків датафрейму**.**

In [3]:

```
df = pd.read_csv("C:\\Users\\Artemii\\Downloads\\train.csv")
print(df.info())
print()
print(df.index)
print(df.dtypes)
print(df.count())
print(df.head(5))
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 891 entries, 0 to 890
Data columns (total 12 columns):
 #   Column       Non-Null Count  Dtype  
---  ------       --------------  -----  
 0   PassengerId  891 non-null    int64  
 1   Survived     891 non-null    int64  
 2   Pclass       891 non-null    int64  
 3   Name         891 non-null    object 
 4   Sex          891 non-null    object 
 5   Age          714 non-null    float64
 6   SibSp        891 non-null    int64  
 7   Parch        891 non-null    int64  
 8   Ticket       891 non-null    object 
 9   Fare         891 non-null    float64
 10  Cabin        204 non-null    object 
 11  Embarked     889 non-null    object 
dtypes: float64(2), int64(5), object(5)
memory usage: 83.7+ KB
None

RangeIndex(start=0, stop=891, step=1)
PassengerId      int64
Survived         int64
Pclass           int64
Name            object
Sex             object
Age            float64
SibSp            int64
Parch            int64
Ticket          object
Fare           float64
Cabin           object
```

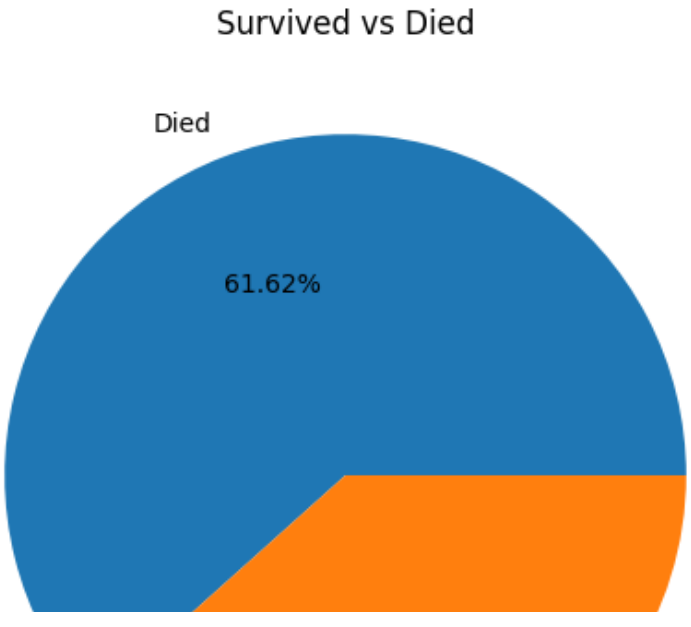
```
Embarked        object
dtype: object
PassengerId    891
Survived       891
Pclass         891
Name           891
Sex            891
Age            714
SibSp          891
Parch          891
Ticket         891
Fare           891
Cabin          204
Embarked       889
dtype: int64
   PassengerId  Survived  Pclass  \
0            1         0       3
1            2         1       1
2            3         1       3
3            4         1       1
4            5         0       3

                                                Name     Sex   Age  SibSp  \
0                            Braund, Mr. Owen Harris    male  22.0      1
1  Cumings, Mrs. John Bradley (Florence Briggs Th...  female  38.0      1
2                             Heikkinen, Miss. Laina  female  26.0      0
3       Futrelle, Mrs. Jacques Heath (Lily May Peel)  female  35.0      1
4                           Allen, Mr. William Henry    male  35.0      0

   Parch            Ticket     Fare Cabin Embarked
0      0         A/5 21171   7.2500   NaN        S
1      0          PC 17599  71.2833   C85        C
2      0  STON/O2. 3101282   7.9250   NaN        S
3      0            113803  53.1000  C123        S
4      0            373450   8.0500   NaN        S
```

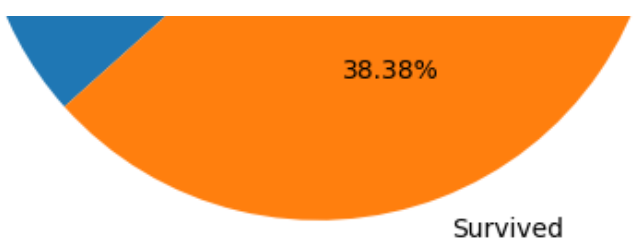
II. Data Visualization.

1. **(10б)** Створіть **pie chart,** який би показував співвідношення вижившиx до загиблих**.**

In [4]:

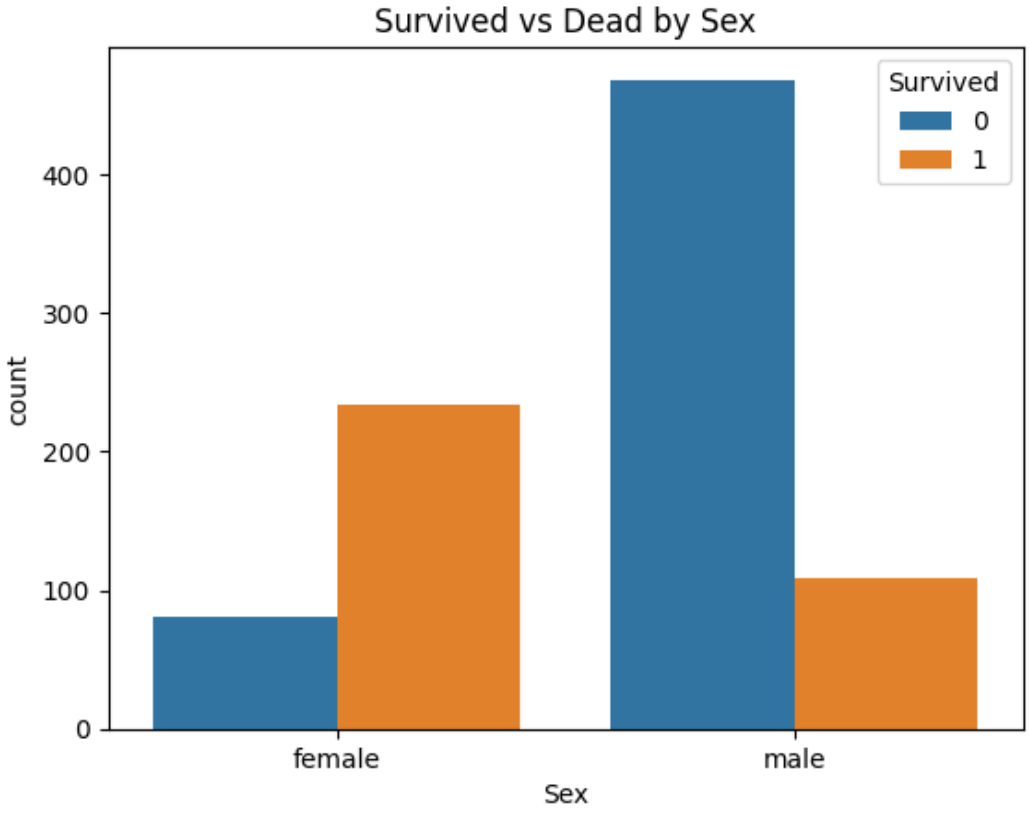
```
survived_counts = df['Survived'].value_counts()

plt.figure(figsize=(6,6))
plt.pie(survived_counts, labels=['Died', 'Survived'], autopct='%1.2f%%')
plt.title('Survived vs Died')
plt.show()
```





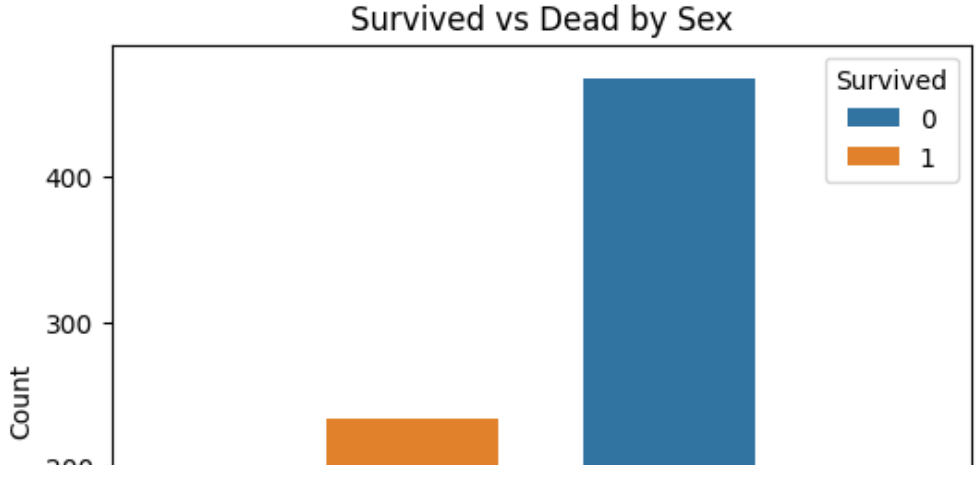
Абсолютна більшість пасажирів загинула**.**

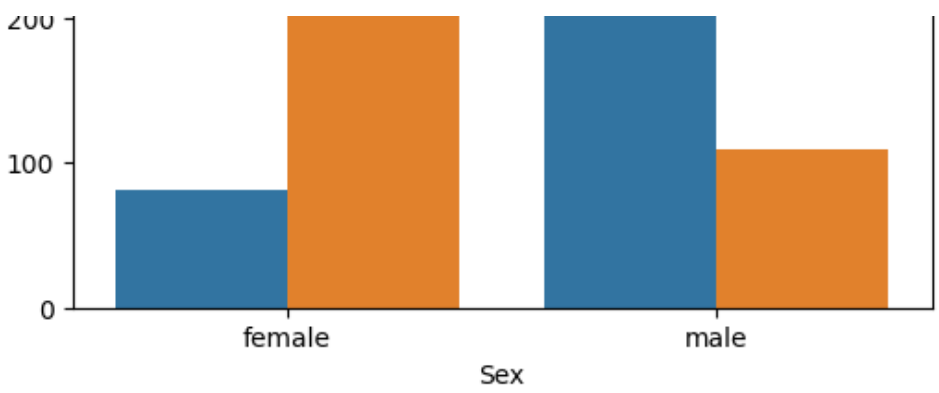
1. **(20**б**)** Створіть **bar chart,** який би показував співвідношення загиблих до вижившив для кожної статі**.** Для **groupby** використовуйте атрибут **as_index=False.**



In [5]:

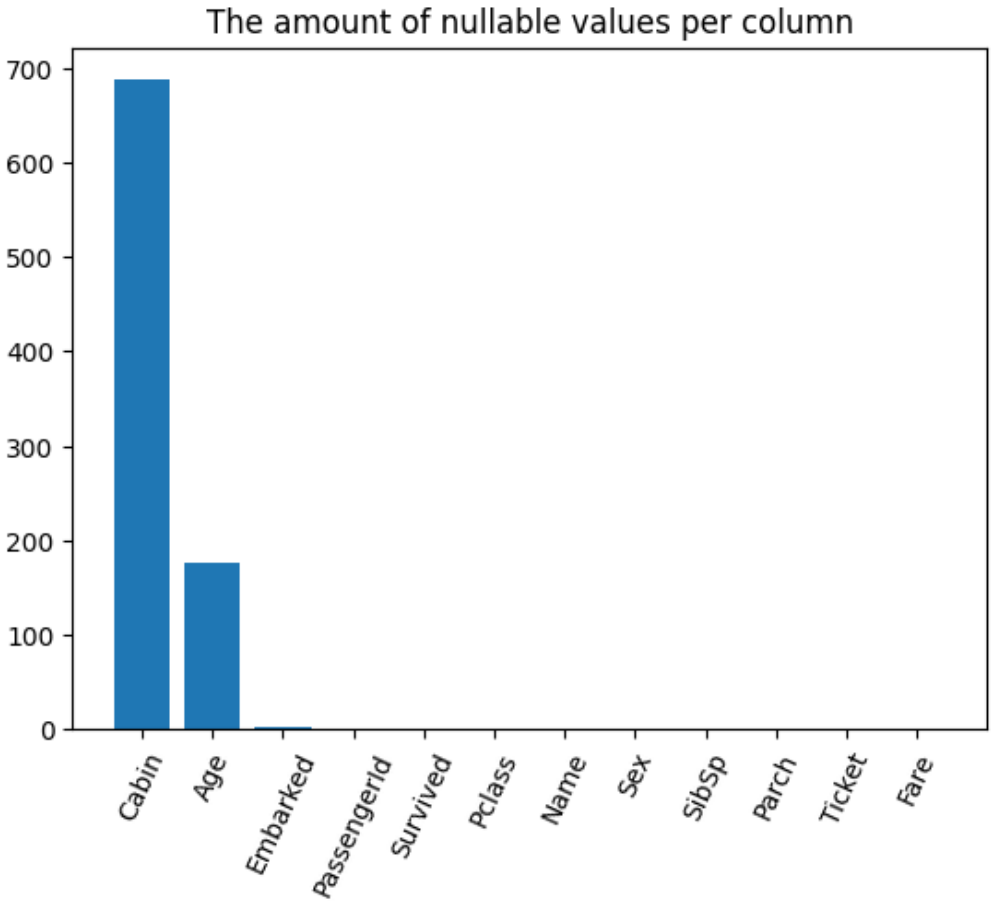
```
grouped = df.groupby(['Sex', 'Survived'], as_index=False).size()

plt.figure(figsize=(6,5))
sns.barplot(x='Sex', y='size', hue='Survived', data=grouped)
plt.title('Survived vs Dead by Sex')
plt.xlabel('Sex')
plt.ylabel('Count')
plt.show()
```





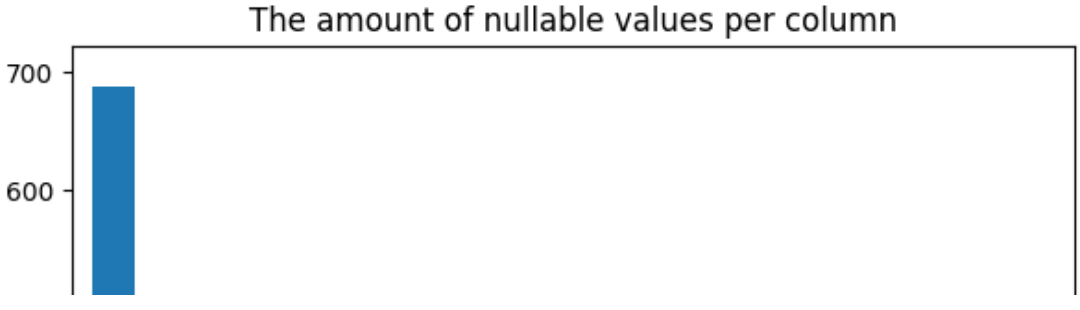
Більшість пасажирів - чоловіки, можна навіть конкретизувати, що чоловіки, які загинули. Майже вдвічі більше було врятовано жінок, аніж чоловіків.

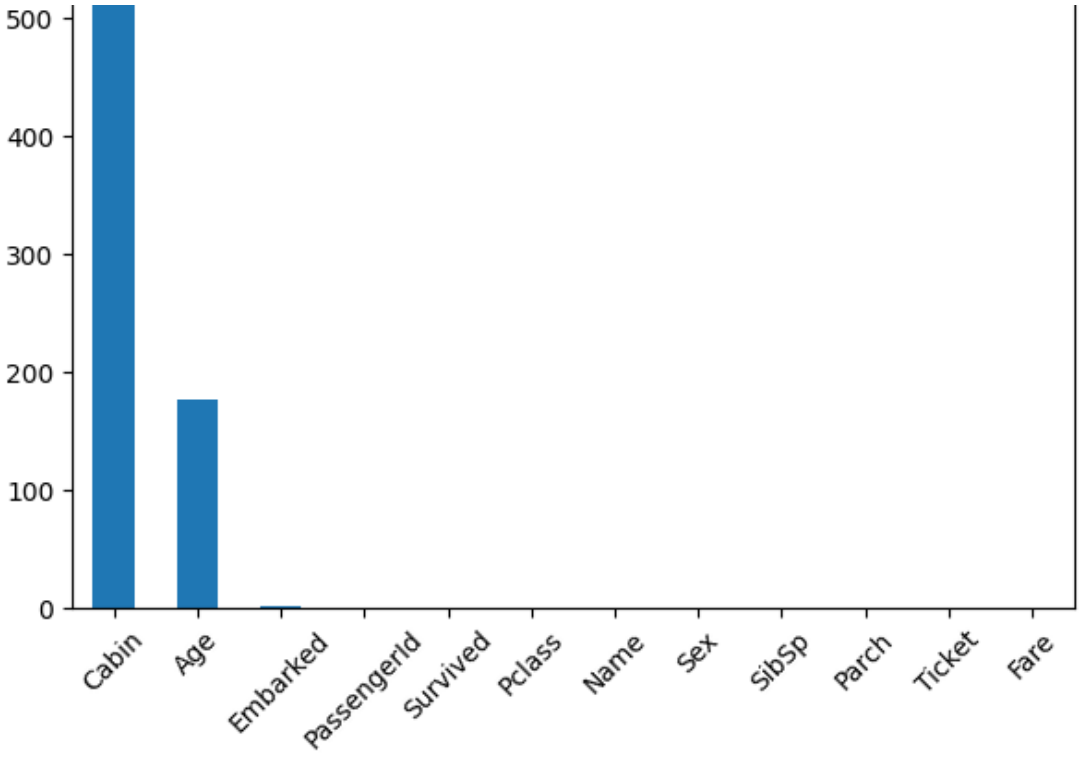
1. **(20б)** Відобразіть кількість пропущених значень в датасеті по кожній із змінних.



In [6]:

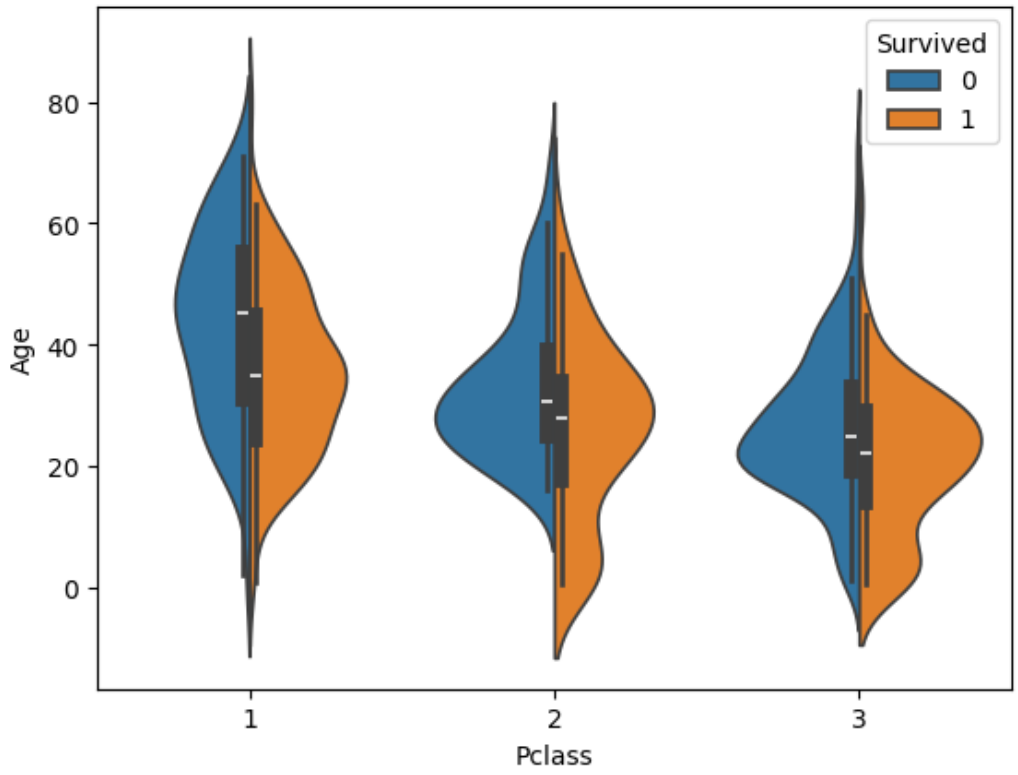
```
missing_values = df.isnull().sum()
missing_values = missing_values.sort_values(ascending=False)

plt.figure(figsize=(7,6))
missing_values.plot(kind='bar')
plt.title('The amount of nullable values per column')
plt.xticks(rotation=45)
plt.show()
```





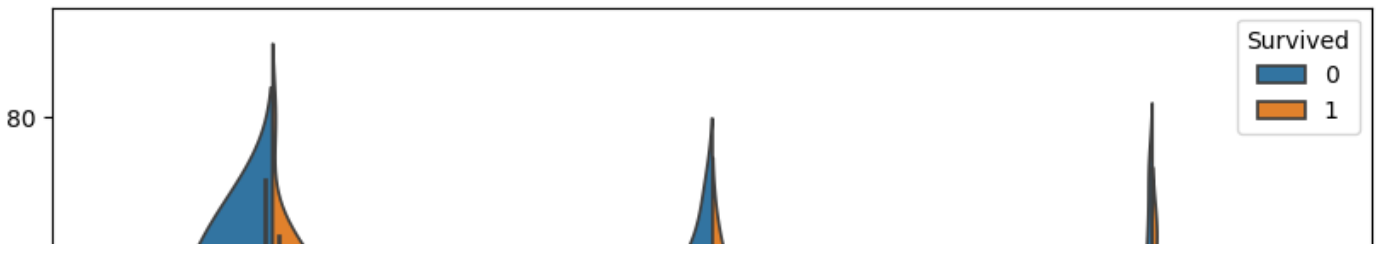
Більшість даних, що була втрачено - про номери кают пасажирів та їхній вік.

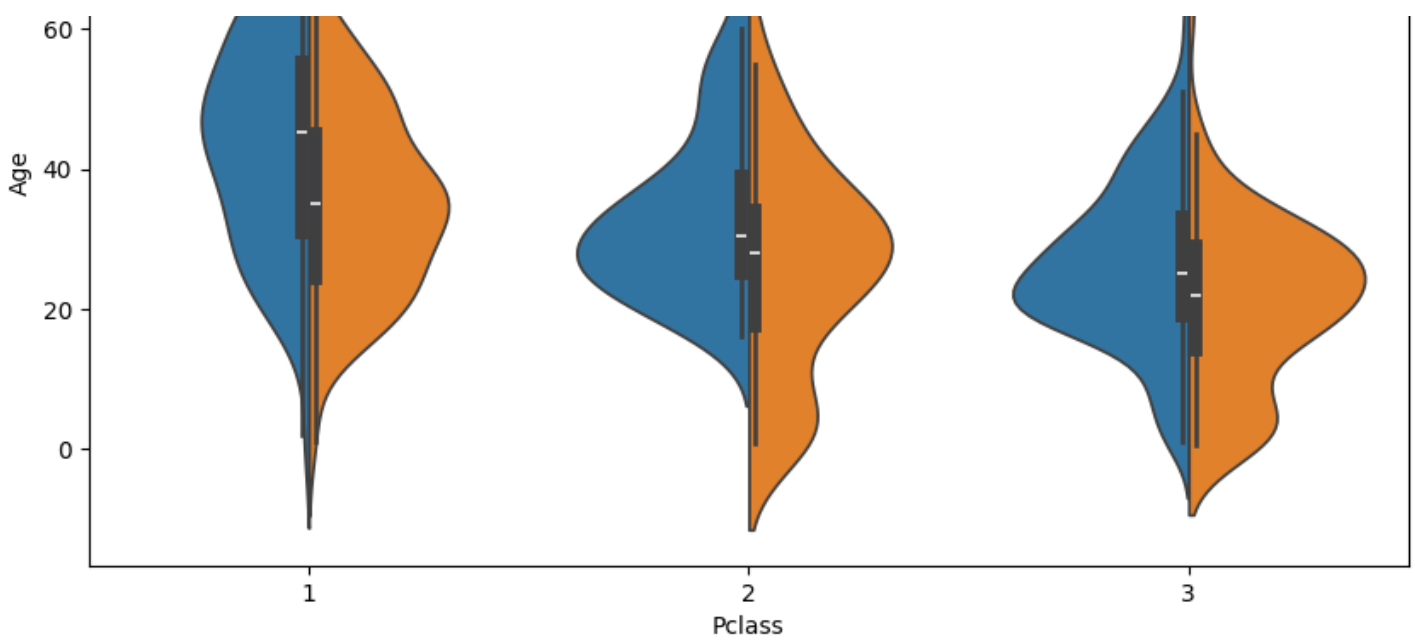
1. **(10б)** Побудуйте наступний графік, дайте йому назву та опишіть, що власне на них відображається.



In [7]:

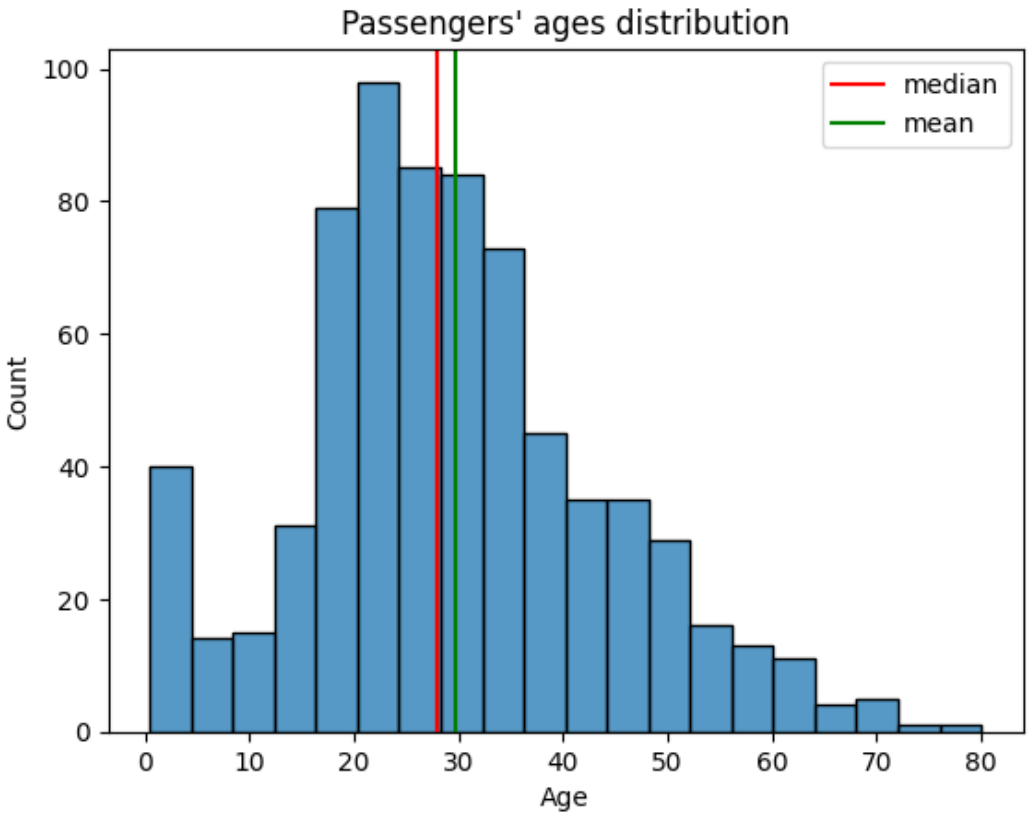
```
plt.figure(figsize=(10,6))
sns.violinplot(x='Pclass', y='Age', hue='Survived', data=df, split=True)
plt.show()
```





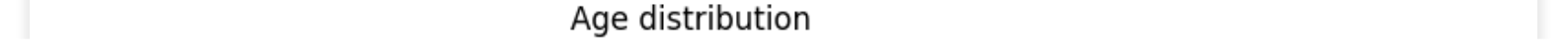
Цей графік показує розподіл пасажирів по класу квитка, їхній вік та чи вижили вони. У **3** класі було більше молодих пасажирів. По першому класу видно, що більше вижили молоді пасажири.

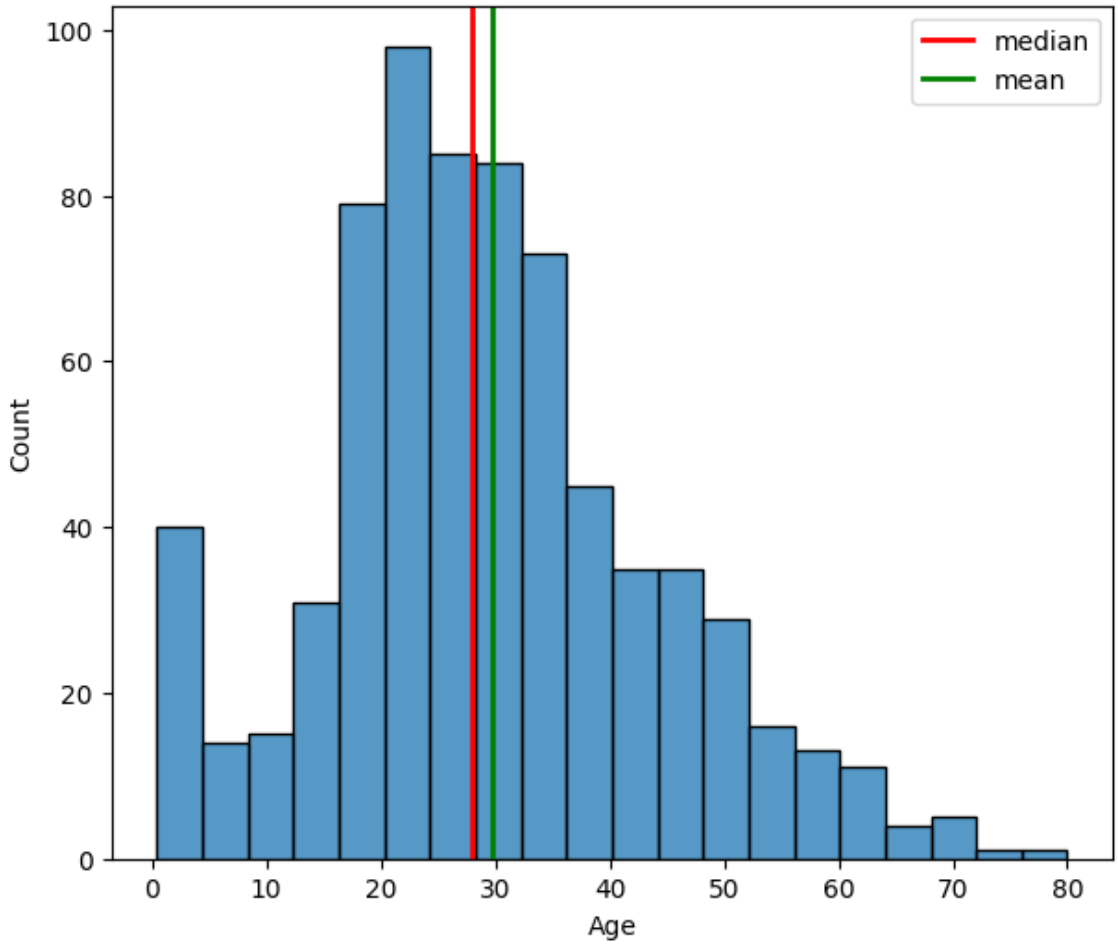
1. **(13б)** Побудуйте графік розподілу частот по віку пасажирів.



In [8]:

```
plt.figure(figsize=(7,6))
sns.histplot(df['Age'], bins=20)

plt.axvline(df['Age'].median(), color='r', linewidth=2, label='median')
plt.axvline(df['Age'].mean(), color='g', linewidth=2, label='mean')

plt.xlabel('Age')
plt.ylabel('Count')
plt.title('Age distribution')
plt.legend()
plt.show()
```





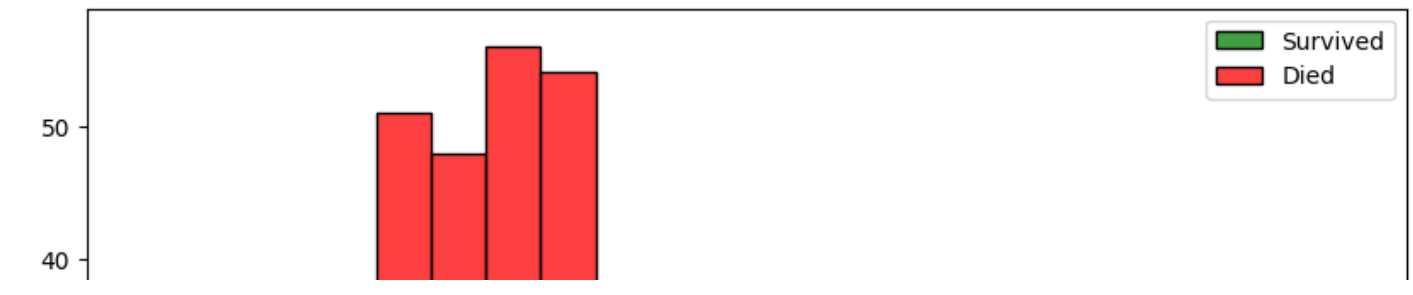
Найбільше було **~20-**річних пасажирів**,** середній вік **~30.**

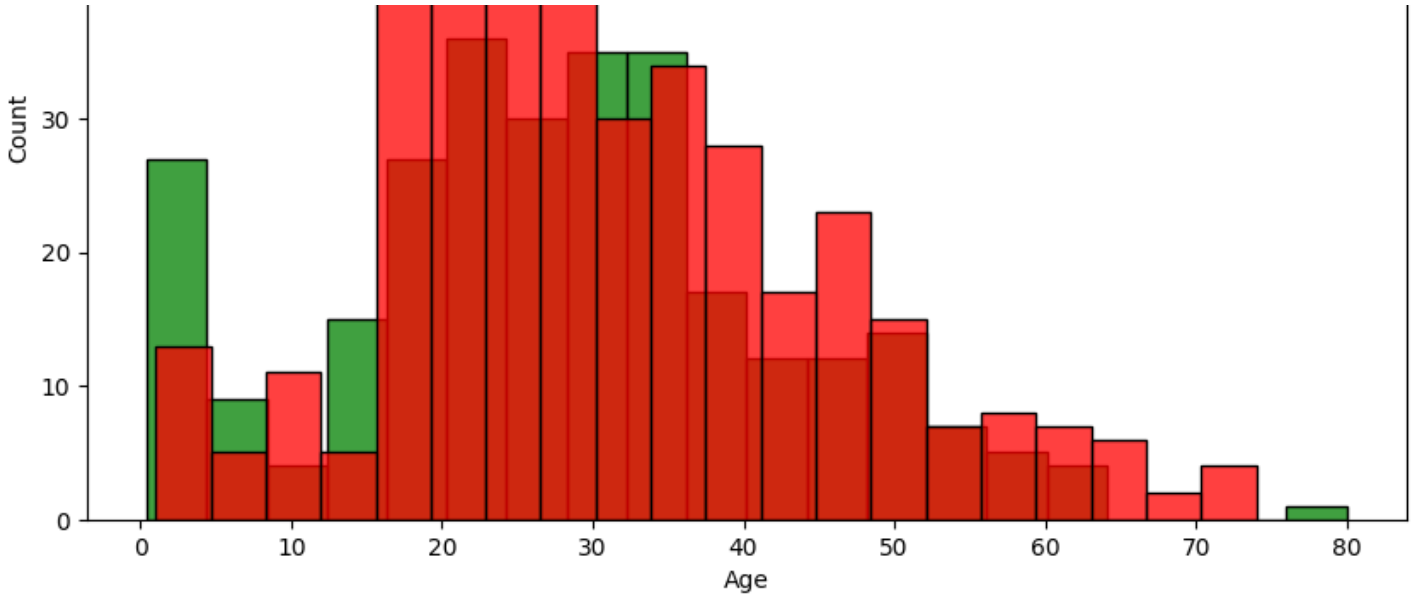
1. **(25**б**)** Поставте бізнес-питання до даних**.** Дайте на нього відповідь за допомогою візуалізацій**.** Прокоментуйте отриману відповідь**.**

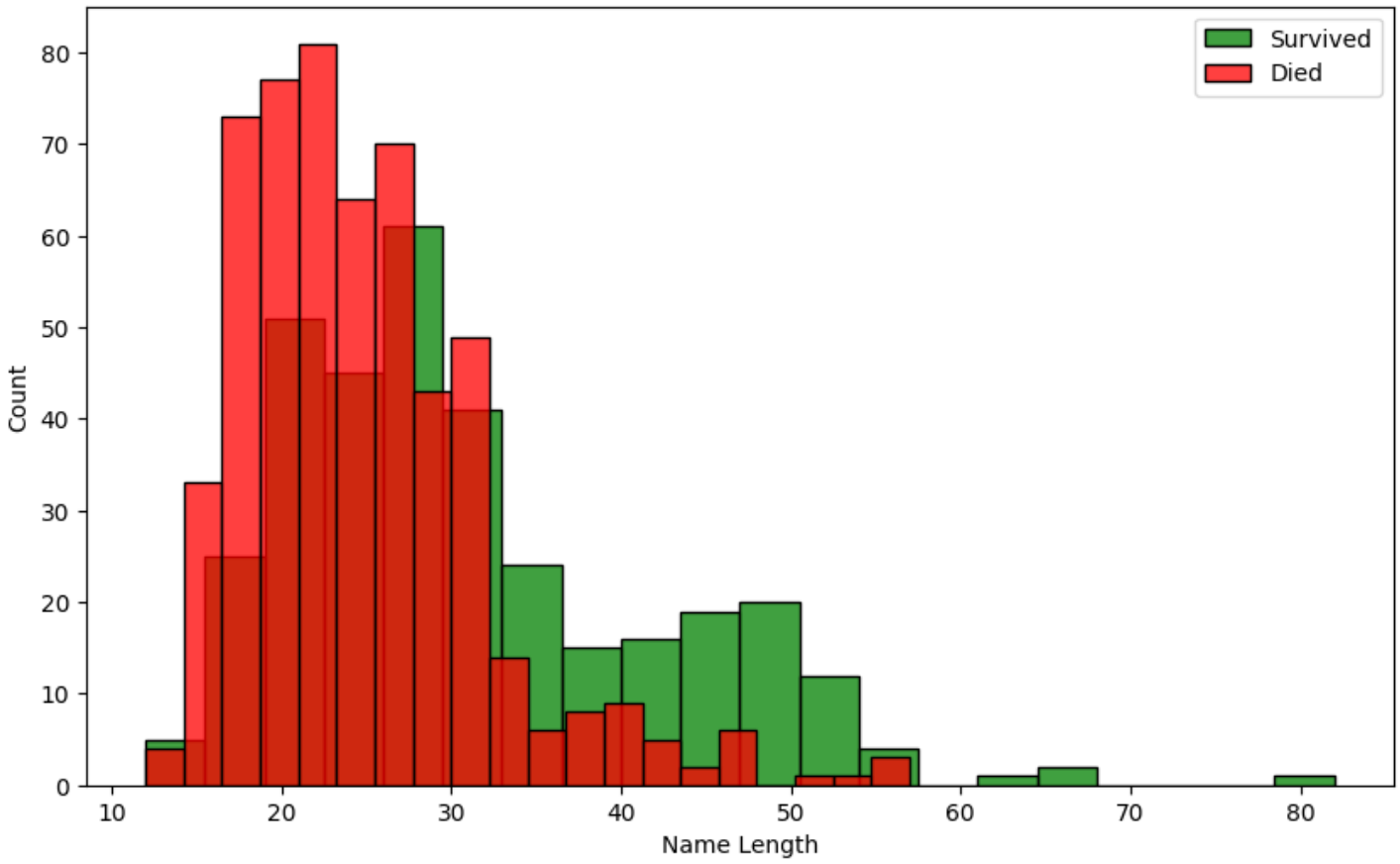
Який зв'язок між різними ознаками та шансом вижити/ хто мав більше шансів вижити**?**

In [18]:

```
plt.figure(figsize=(10,6))
sns.histplot(df[df['Survived'] == 1]['Age'], bins=20, color='green', label='Survived')
sns.histplot(df[df['Survived'] == 0]['Age'], bins=20, color='red', label='Died')
plt.xlabel('Age')
plt.ylabel('Count')
plt.legend()
plt.show()


df['Name_length'] = df['Name'].apply(len)
plt.figure(figsize=(10,6))
sns.histplot(df[df['Survived'] == 1]['Name_length'], bins=20, color='green', label='Surv
ived')
sns.histplot(df[df['Survived'] == 0]['Name_length'], bins=20, color='red', label='Died')
plt.xlabel('Name Length')
plt.ylabel('Count')
plt.legend()
plt.show()
```

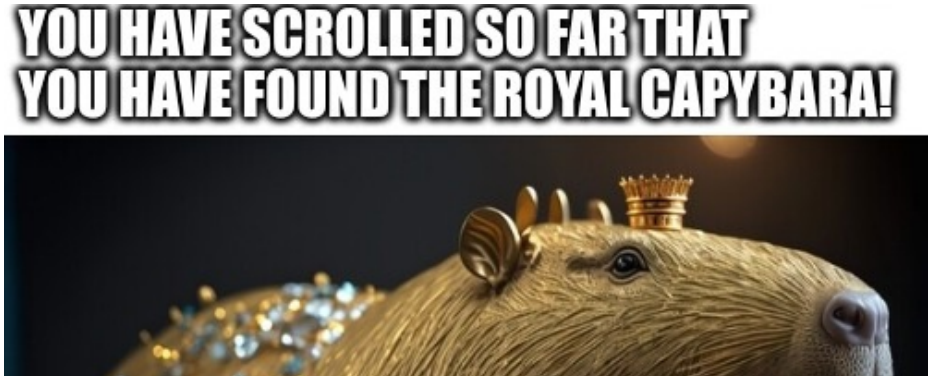






Найбільше шансів вижити в молодих пасажирів до **~15** та в тих, хто мав довге ім'я, напевно, ці пасажири були заможнішими.

Вітаю! Ви велика(ий) молодець, що впоралась(вся). Похваліть себе та побалуйте чимось приємним. Я Вами пишаюся.




YOU MAY TAKE HIM AS YOUR REWARD!
imgflip.com