In [27]:
```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
%matplotlib inline
```

In [28]:
```
df = pd.read_csv('./titanic.csv')
df.head()
```

Out[28]:

	survived	pclass	sex	age	sibsp	parch	fare	embarked	class	who	adult_m
0	0	3	male	22.0	1	0	7.2500	S	Third	man	T
1	1	1	female	38.0	1	0	71.2833	C	First	woman	Fa
2	1	3	female	26.0	0	0	7.9250	S	Third	woman	Fa
3	1	1	female	35.0	1	0	53.1000	S	First	woman	Fa
4	0	3	male	35.0	0	0	8.0500	S	Third	man	T

In [29]:
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 891 entries, 0 to 890
Data columns (total 15 columns):
 #   Column       Non-Null Count  Dtype
---  ------       --------------  -----
 0   survived     891 non-null    int64
 1   pclass       891 non-null    int64
 2   sex          891 non-null    object
 3   age          714 non-null    float64
 4   sibsp        891 non-null    int64
 5   parch        891 non-null    int64
 6   fare         891 non-null    float64
 7   embarked     889 non-null    object
 8   class        891 non-null    object
 9   who          891 non-null    object
 10  adult_male   891 non-null    bool
 11  deck         203 non-null    object
 12  embark_town  889 non-null    object
 13  alive        891 non-null    object
 14  alone        891 non-null    bool
dtypes: bool(2), float64(2), int64(4), object(7)
memory usage: 92.4+ KB
```

In [30]:
```
df.describe()
```

Out[30]:

	survived	pclass	age	sibsp	parch	fare
count	891.000000	891.000000	714.000000	891.000000	891.000000	891.000000
mean	0.383838	2.308642	29.699118	0.523008	0.381594	32.204208
std	0.486592	0.836071	14.526497	1.102743	0.806057	49.693429
min	0.000000	1.000000	0.420000	0.000000	0.000000	0.000000
25%	0.000000	2.000000	20.125000	0.000000	0.000000	7.910400
50%	0.000000	3.000000	28.000000	0.000000	0.000000	14.454200
75%	1.000000	3.000000	38.000000	1.000000	0.000000	31.000000
max	1.000000	3.000000	80.000000	8.000000	6.000000	512.329200

In [31]:
```
df.isnull().sum()
```

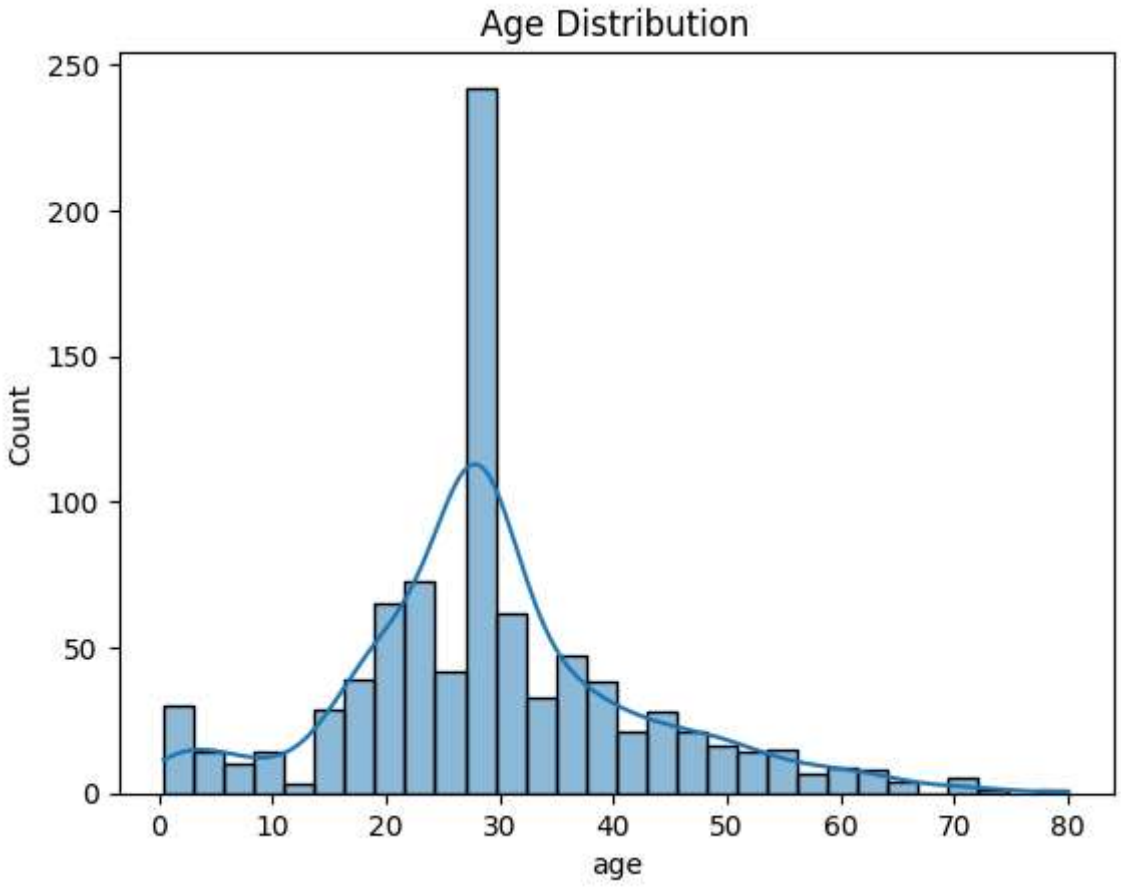
Out[31]:
```
survived         0
pclass           0
sex              0
age            177
sibsp            0
parch            0
fare             0
embarked         2
class            0
who              0
adult_male       0
deck           688
embark_town      2
alive            0
alone            0
dtype: int64
```

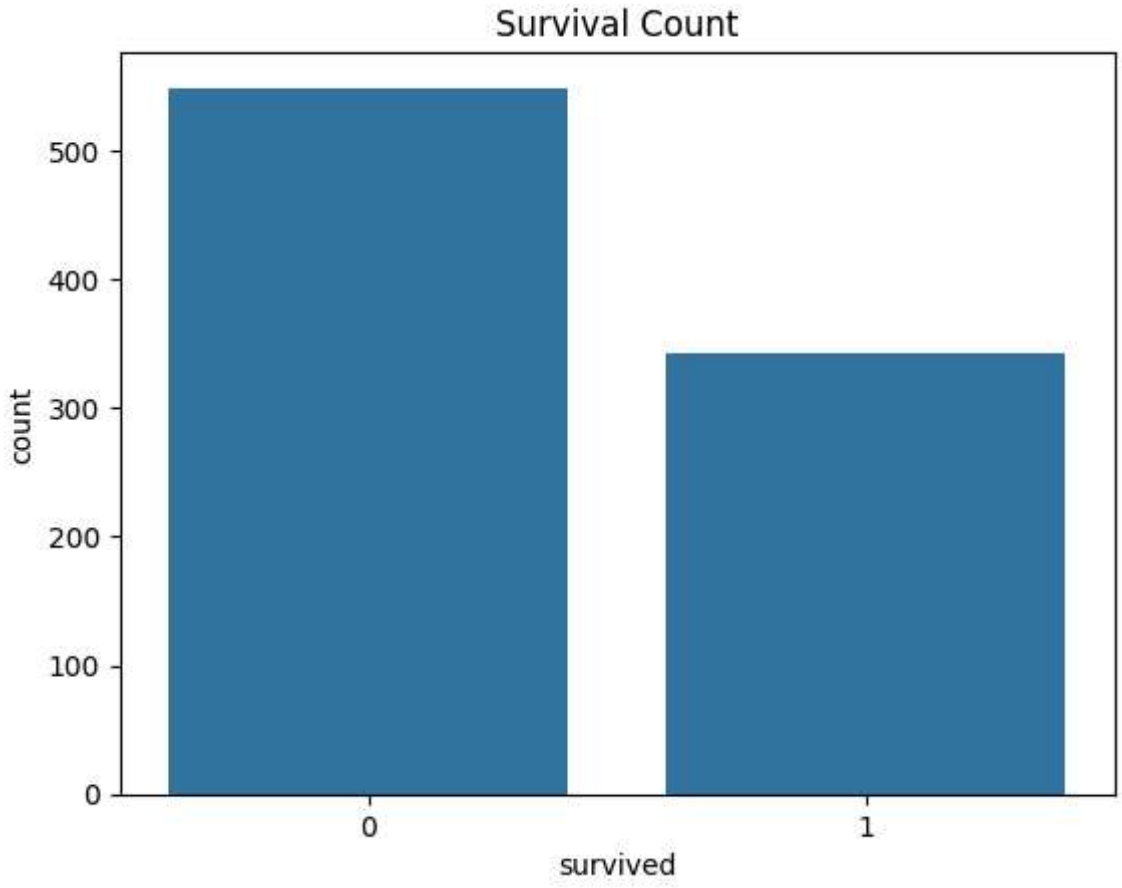
In [32]:
```
df['age'] = df['age'].fillna(df['age'].median())
df['embarked'] = df['embarked'].fillna(df['embarked'].mode()[0])

df = df.drop('deck', axis=1)
df = df.drop('embark_town', axis=1)
```

In [33]:
```
df.isnull().sum()
```

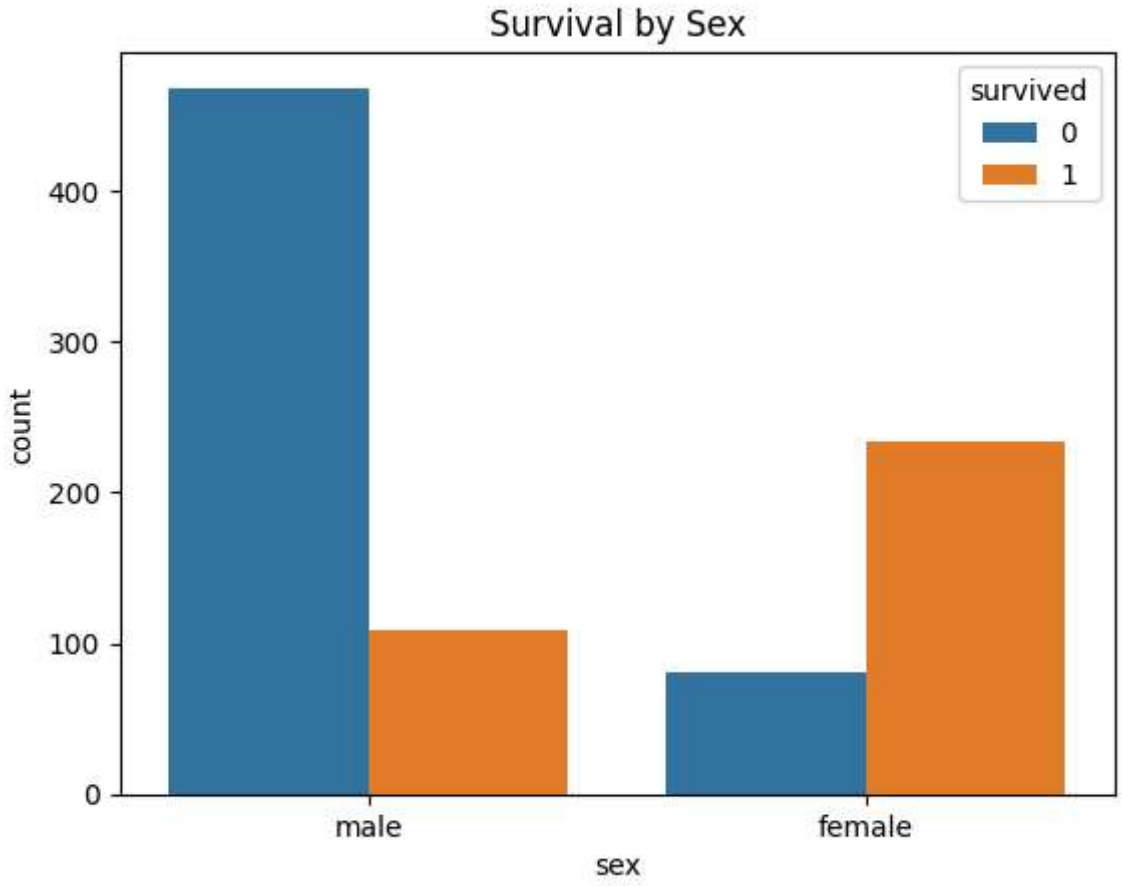
Out[33]:
```
survived      0
pclass        0
sex           0
age           0
sibsp         0
parch         0
fare          0
embarked      0
class         0
who           0
adult_male    0
alive         0
alone         0
dtype: int64
```

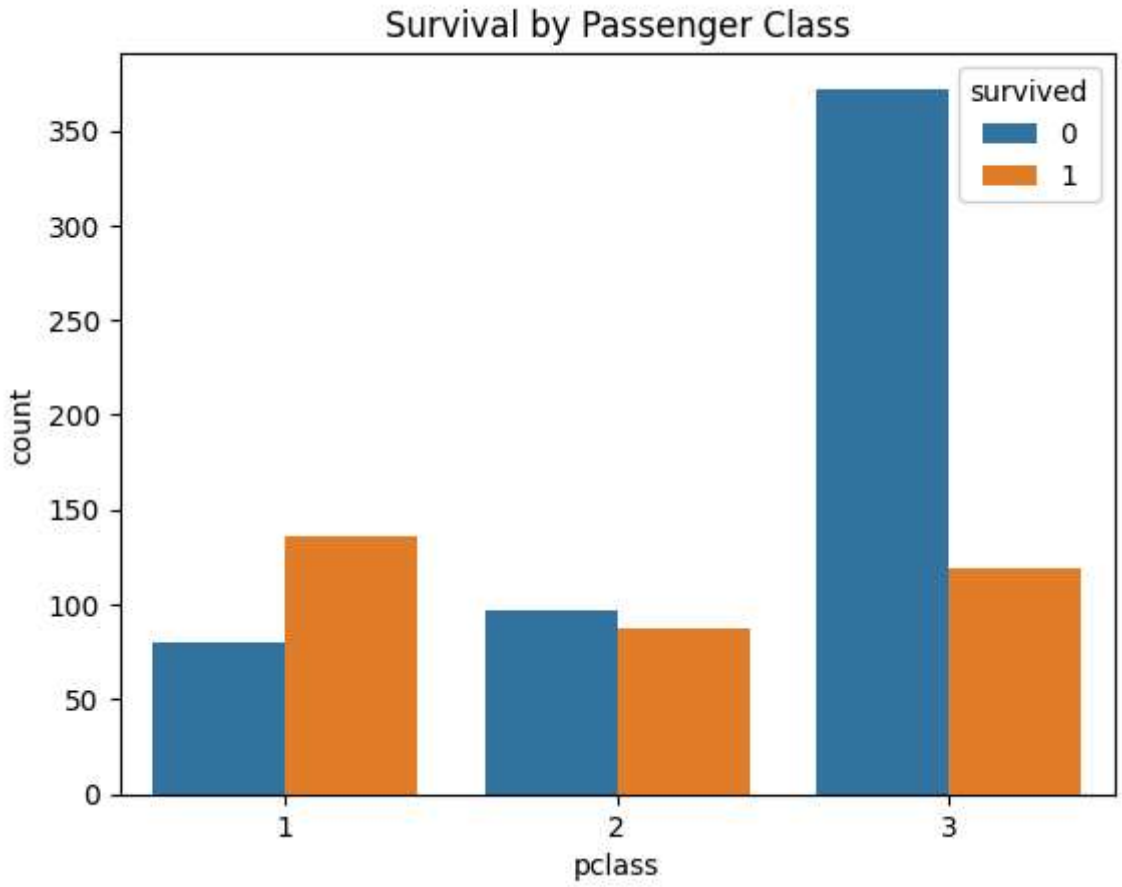
In [34]:
```
sns.histplot(df['age'], kde=True)
plt.title('Age Distribution')
plt.show()

sns.countplot(x='survived', data=df)
plt.title('Survival Count')
plt.show()
```



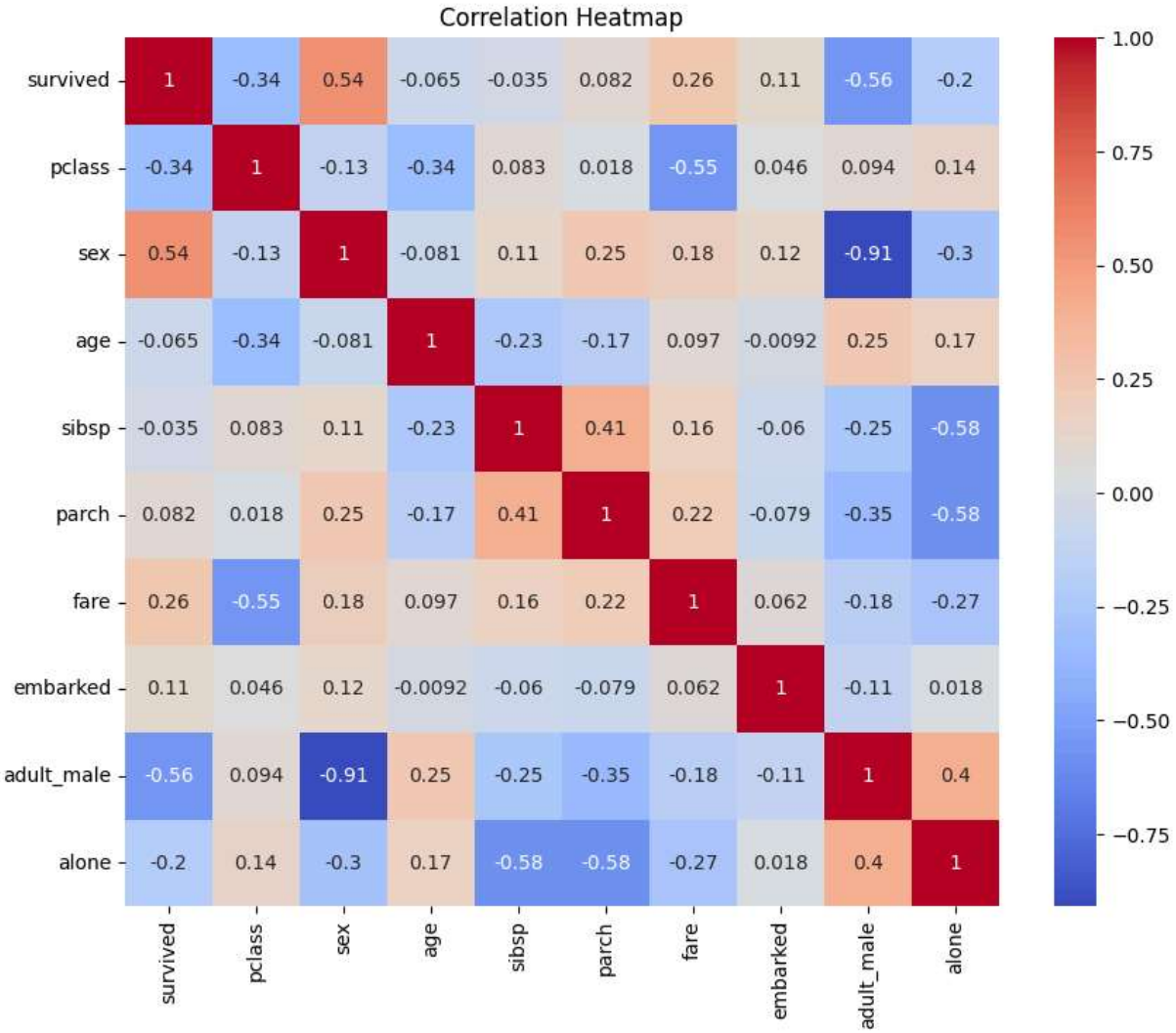


In [35]:
```
sns.countplot(x='sex', hue='survived', data=df)
plt.title('Survival by Sex')
plt.show()

sns.countplot(x='pclass', hue='survived', data=df)
plt.title('Survival by Passenger Class')
plt.show()
```

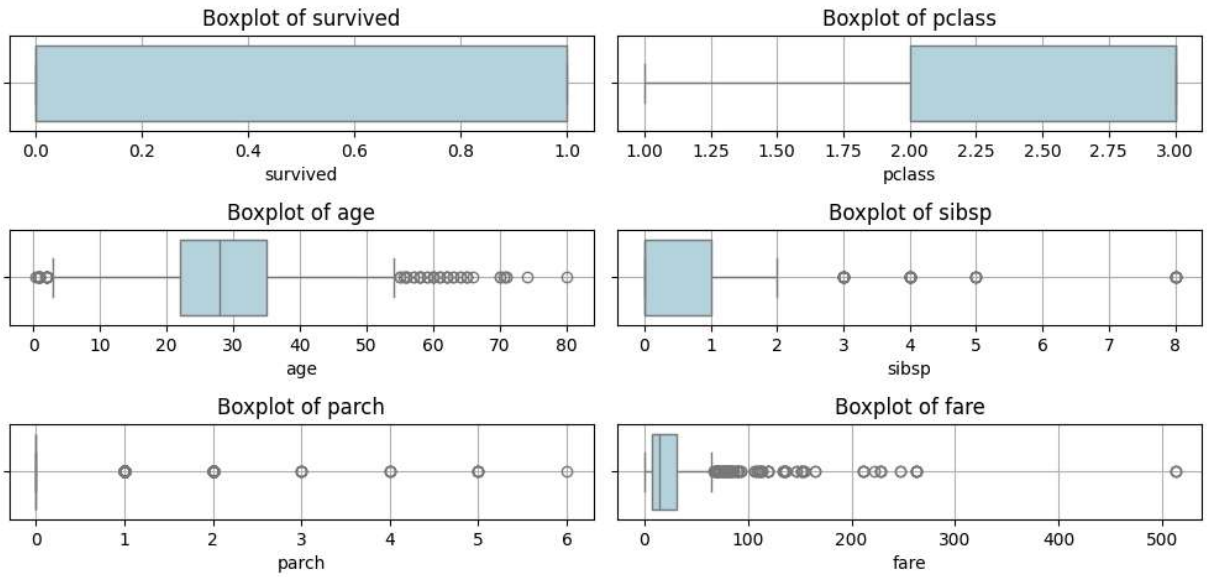

Survival by Sex
survived
0
1
400
300
200
100
0
count
male
female
sex


Survival by Passenger Class
survived
0
1
350
300
250
200
150
100
50
0
count
1
2
3
pclass

In [36]:
```
df_corr = df.copy()
df_corr['sex'] = df_corr['sex'].map({'male':0, 'female':1})
df_corr['embarked'] = df_corr['embarked'].map({'S':0, 'C':1, 'Q':2})

plt.figure(figsize=(10,8))
sns.heatmap(df_corr.corr(numeric_only=True), annot=True, cmap='coolwarm')
plt.title('Correlation Heatmap')
plt.show()
```



In [37]:
```
numeric_columns = df.select_dtypes(include=['number']).columns

plt.figure(figsize=(10, 6))

for i, column in enumerate(numeric_columns):
    plt.subplot(4, 2, i + 1)
    sns.boxplot(x=df[column], color='lightblue')
    plt.title(f'Boxplot of {column}')
    plt.xlabel(column)
    plt.grid(True)

plt.tight_layout()
plt.show()
```



In [38]:
```
from ydata_profiling import ProfileReport

profile = ProfileReport(df, title='Titanic Dataset EDA')
```

In [39]:
```
import os
os.makedirs('../reports', exist_ok=True)

profile.to_file('../reports/eda_titanic_report.html')
```

```
Summarize dataset:   0%|          | 0/5 [00:00<?, ?it/s]
100%|██████████| 13/13 [00:00<00:00, 709.23it/s]
Generate report structure:   0%|          | 0/1 [00:00<?, ?it/s]
Render HTML:   0%|          | 0/1 [00:00<?, ?it/s]
Export report to file:   0%|          | 0/1 [00:00<?, ?it/s]
```

In [ ]: