

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

```
In [2]: %matplotlib inline
```

```
In [3]: df = pd.read_csv("train.csv")
```

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

```
Out[4]:
```

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th... Heikkinen, Miss. Laina	female	38.0	1	0	PC 17599 STON/O2. 3101282	71.2833 7.9250
2	3	1	3	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	26.0	0	0	113803	53.1000
4	5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.0500



```
In [6]: df.shape
```

```
Out[6]: (891, 12)
```

```
In [7]: df.columns
```

```
Out[7]: Index(['PassengerId', 'Survived', 'Pclass', 'Name', 'Sex', 'Age', 'SibSp',
   'Parch', 'Ticket', 'Fare', 'Cabin', 'Embarked'],
   dtype='object')
```

```
In [8]: df.info()
```

```
<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
```

In [9]: `df.describe()`

	PassengerId	Survived	Pclass	Age	SibSp	Parch	Fare
<b>count</b>	891.000000	891.000000	891.000000	714.000000	891.000000	891.000000	891.000000
<b>mean</b>	446.000000	0.383838	2.308642	29.699118	0.523008	0.381594	32.204208
<b>std</b>	257.353842	0.486592	0.836071	14.526497	1.102743	0.806057	49.693429
<b>min</b>	1.000000	0.000000	1.000000	0.420000	0.000000	0.000000	0.000000
<b>25%</b>	223.500000	0.000000	2.000000	20.125000	0.000000	0.000000	7.910400
<b>50%</b>	446.000000	0.000000	3.000000	28.000000	0.000000	0.000000	14.454200
<b>75%</b>	668.500000	1.000000	3.000000	38.000000	1.000000	0.000000	31.000000
<b>max</b>	891.000000	1.000000	3.000000	80.000000	8.000000	6.000000	512.329200



In [10]: `df.isnull().sum()`

```
Out[10]: PassengerId      0
Survived        0
Pclass          0
Name            0
Sex             0
Age           177
SibSp          0
Parch          0
Ticket         0
Fare           0
Cabin         687
Embarked       2
dtype: int64
```

```
In [13]: df['Age'] = df['Age'].fillna(df['Age'].median())
```

```
In [14]: df['Embarked'] = df['Embarked'].fillna(df['Embarked'].mode()[0])
```

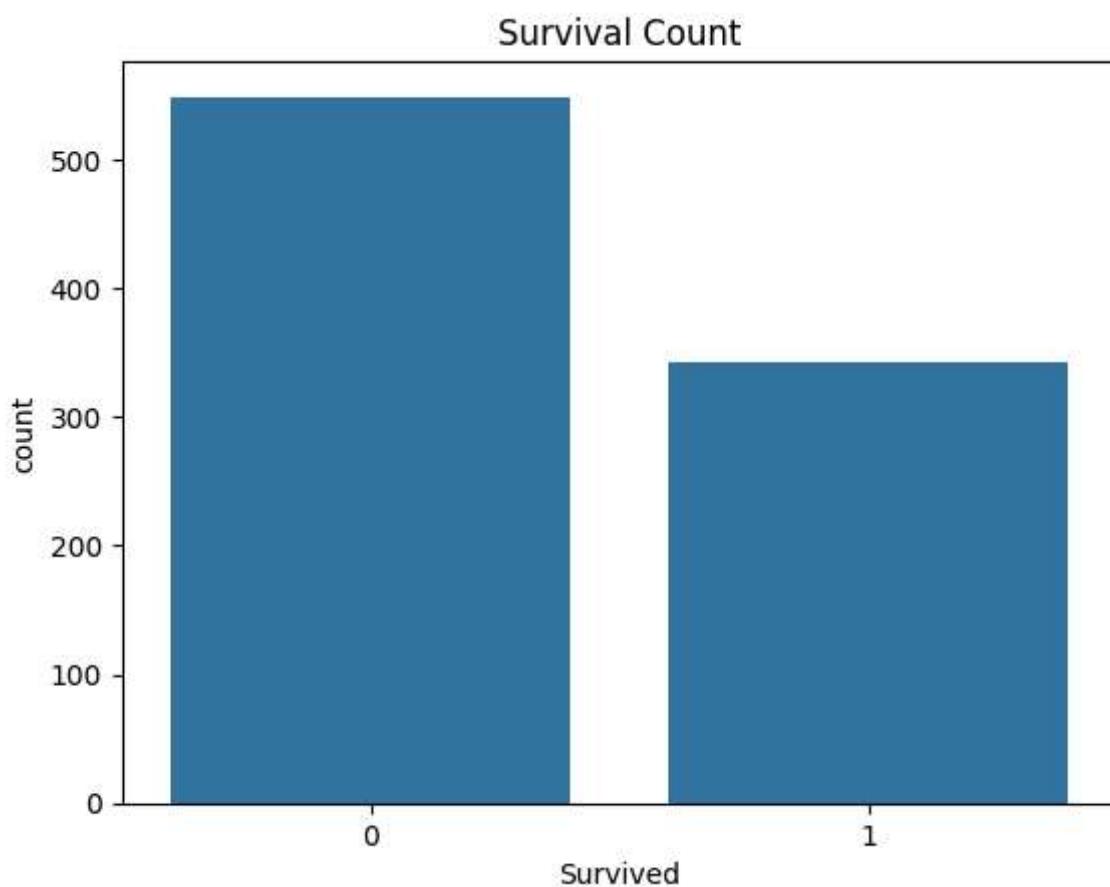
```
In [15]: df.drop(columns=['Cabin'], inplace=True)
```

```
In [16]: df.isnull().sum()
```

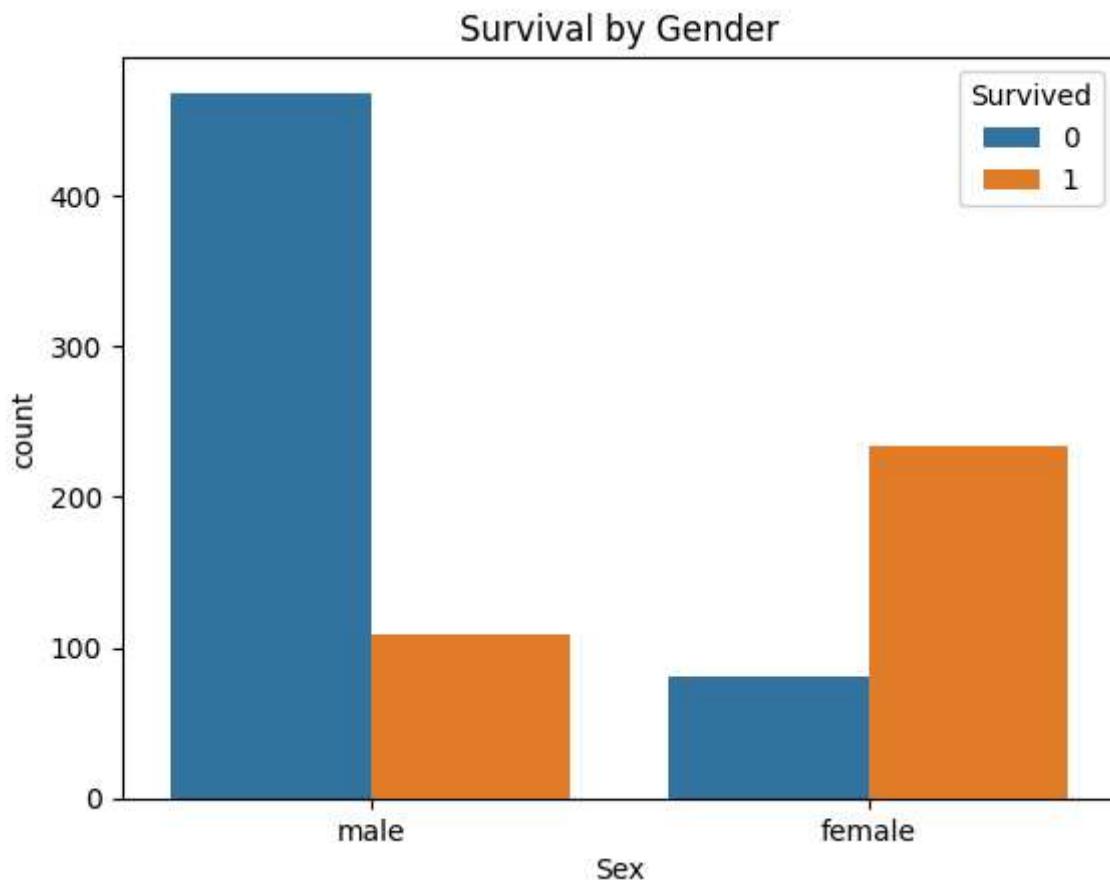
```
Out[16]: PassengerId      0  
Survived          0  
Pclass            0  
Name              0  
Sex               0  
Age               0  
SibSp             0  
Parch             0  
Ticket            0  
Fare              0  
Embarked          0  
dtype: int64
```

```
In [17]: df['FamilySize'] = df['SibSp'] + df['Parch'] + 1
```

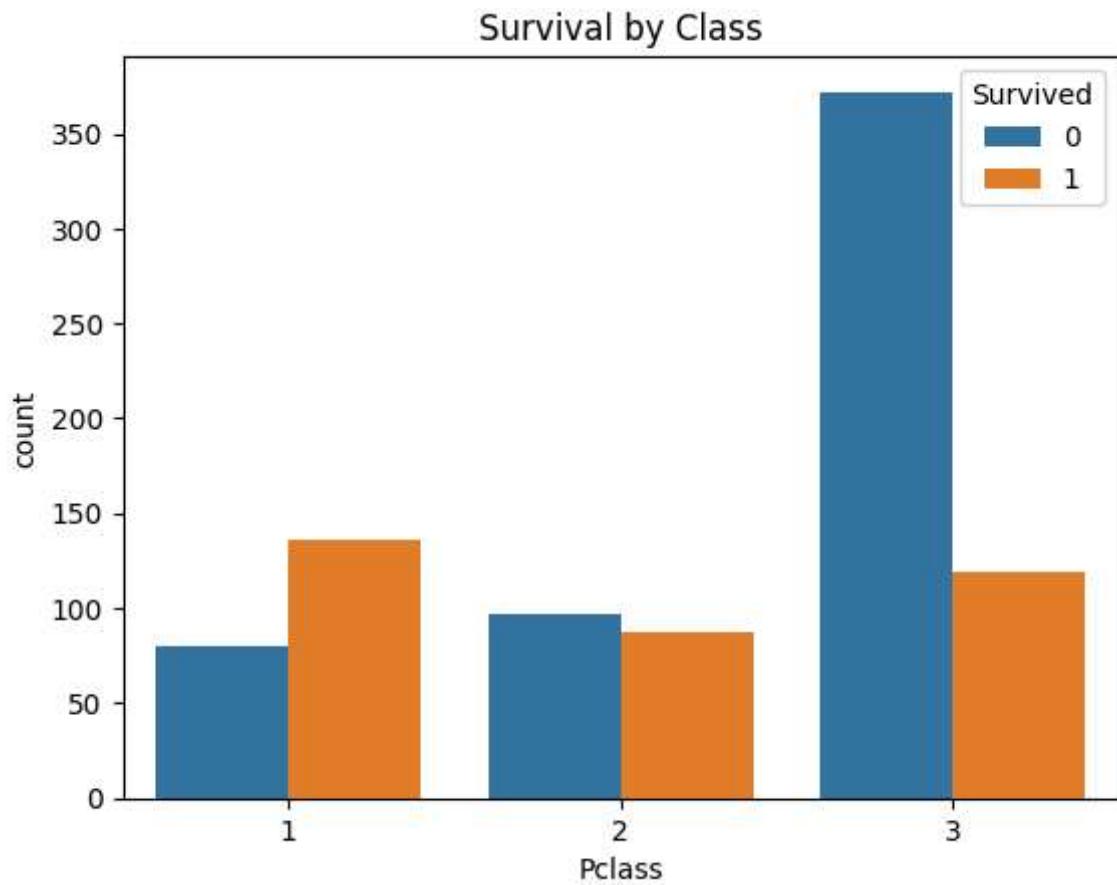
```
In [20]: sns.countplot(x='Survived', data=df)  
plt.title('Survival Count')  
plt.show()
```



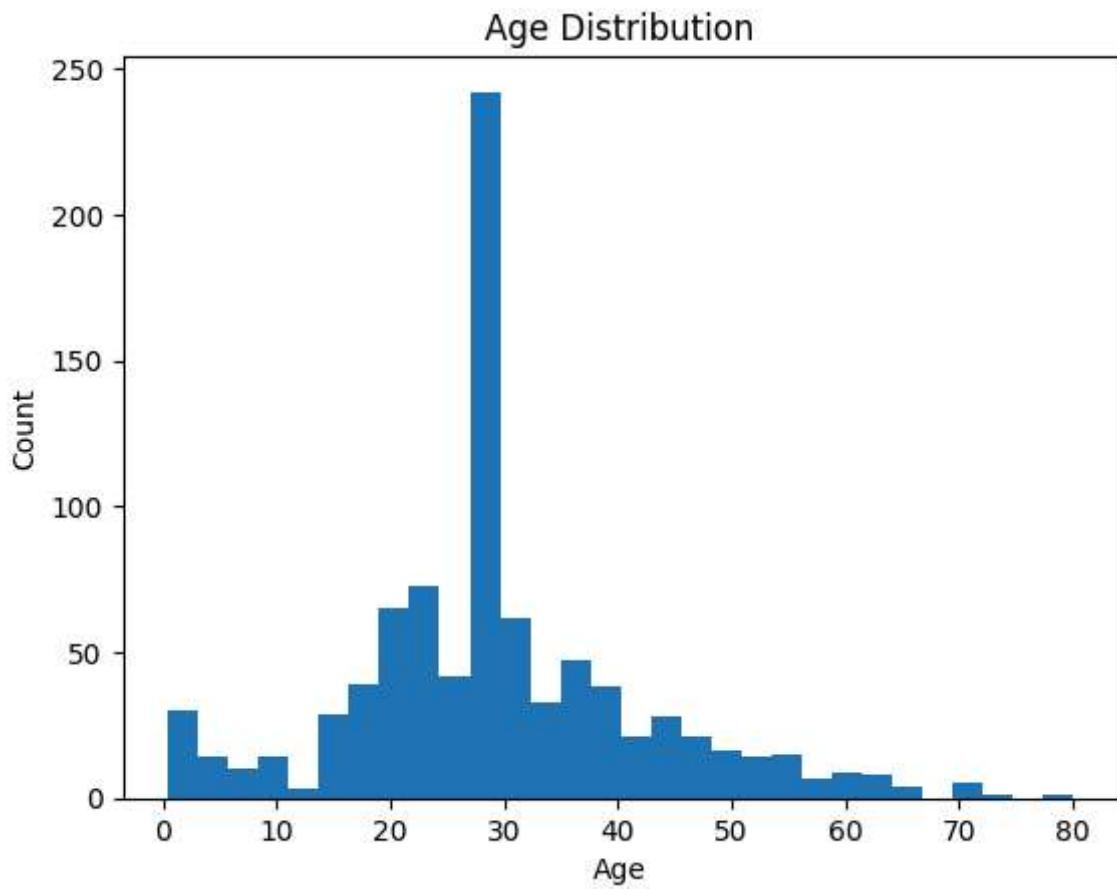
```
In [21]: sns.countplot(x='Sex', hue='Survived', data=df)
plt.title('Survival by Gender')
plt.show()
```



```
In [22]: sns.countplot(x='Pclass', hue='Survived', data=df)
plt.title('Survival by Class')
plt.show()
```

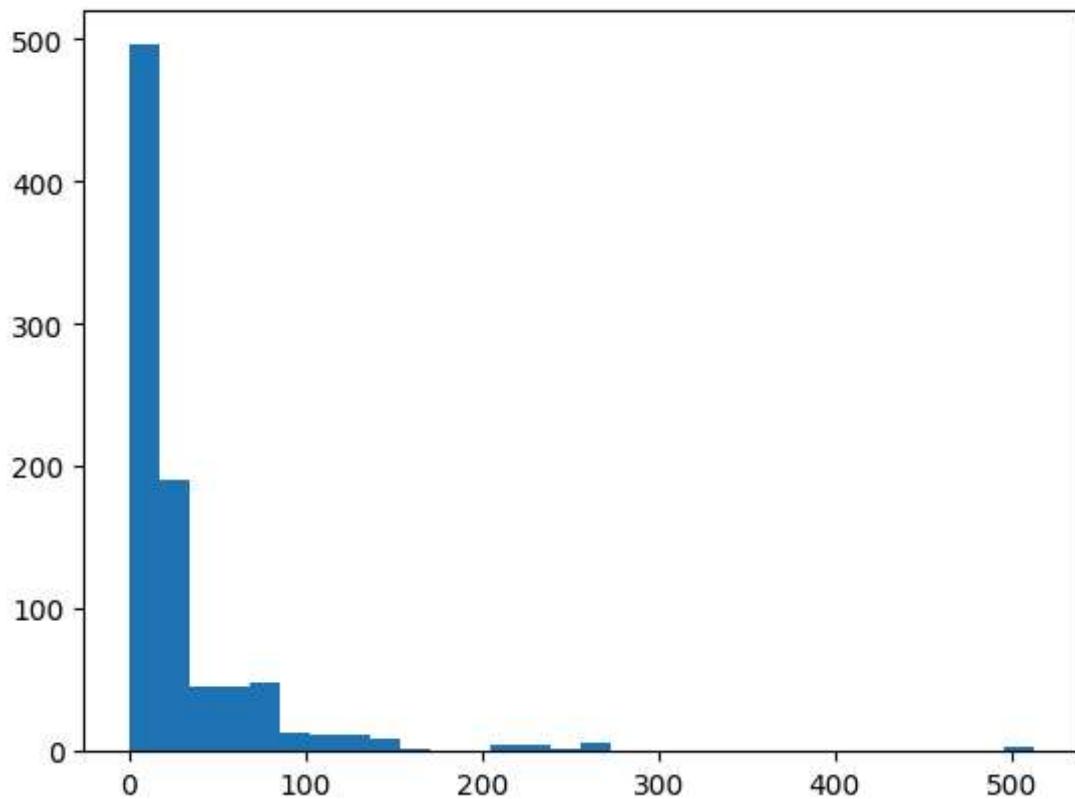


```
In [23]: plt.hist(df['Age'], bins=30)
plt.title("Age Distribution")
plt.xlabel("Age")
plt.ylabel("Count")
plt.show()
```

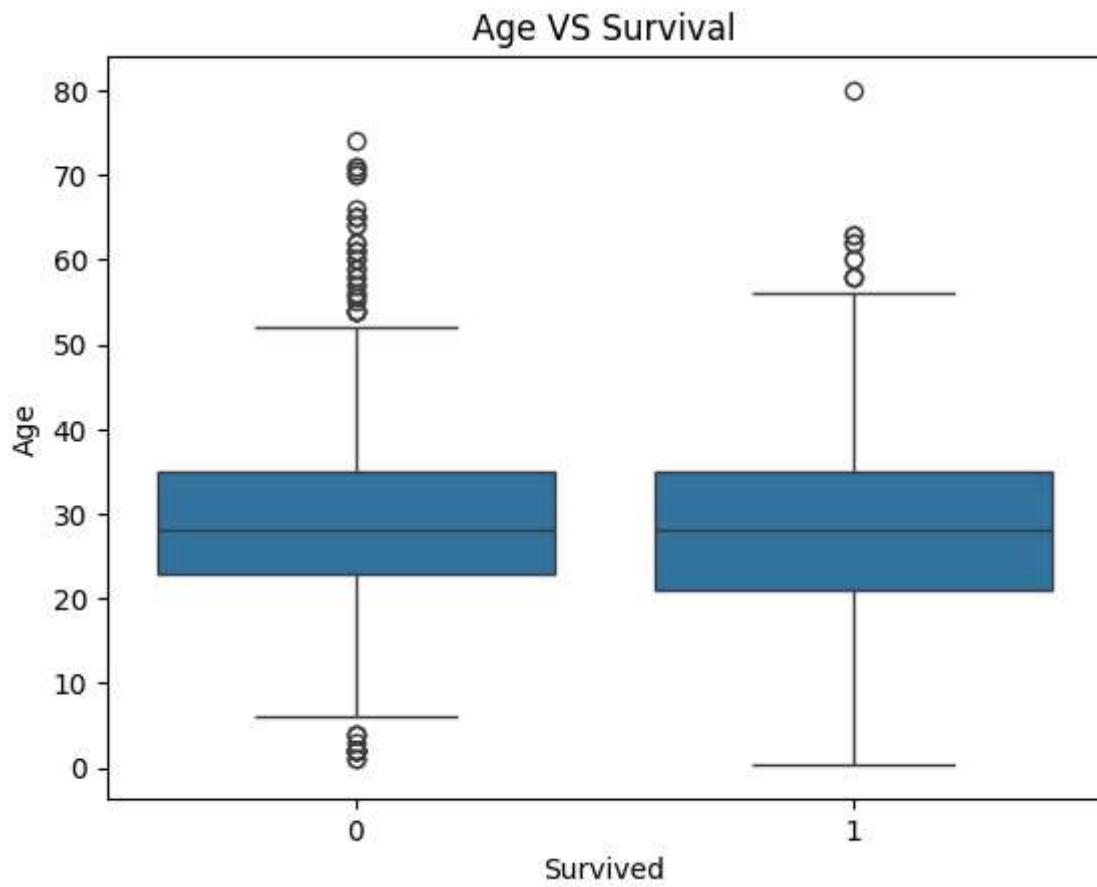


```
In [24]: plt.hist(df['Fare'], bins=30)
plt.title("Fare Distribution")
plt.show()
```

### Fare Distribution

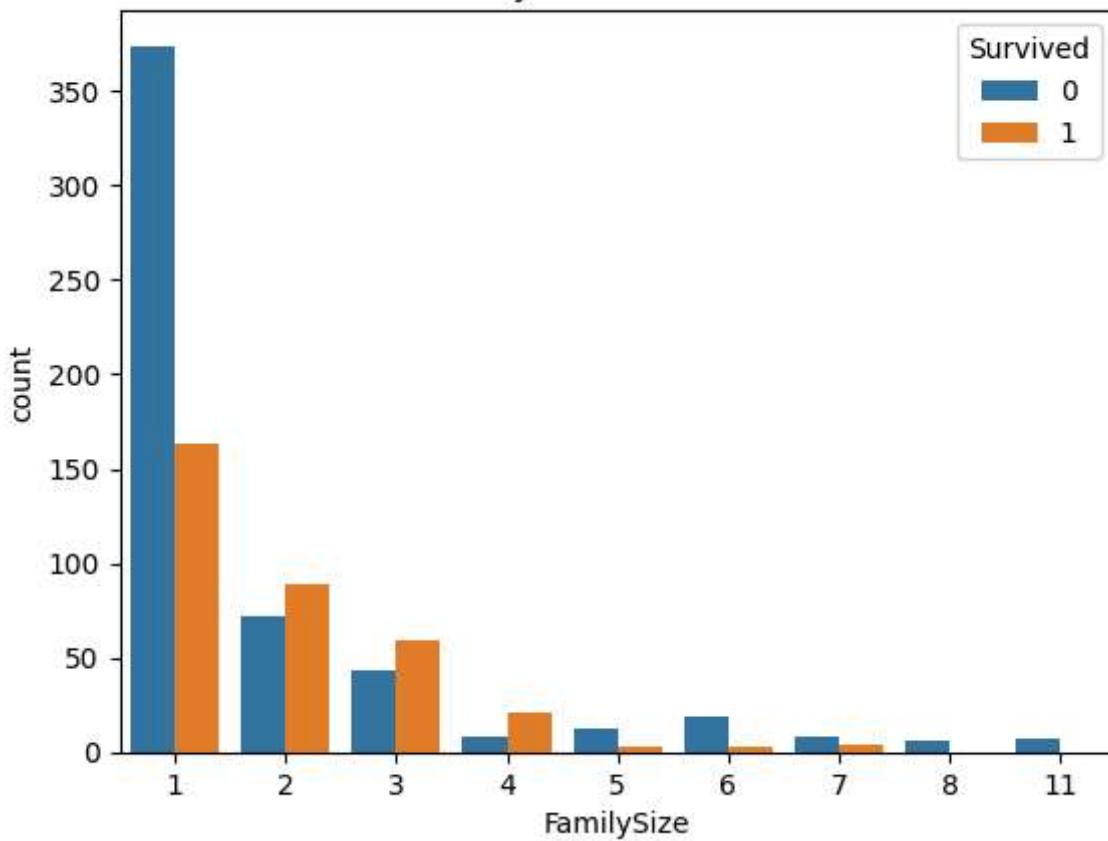


```
In [25]: sns.boxplot(x='Survived', y='Age', data=df)
plt.title("Age VS Survival")
plt.show()
```

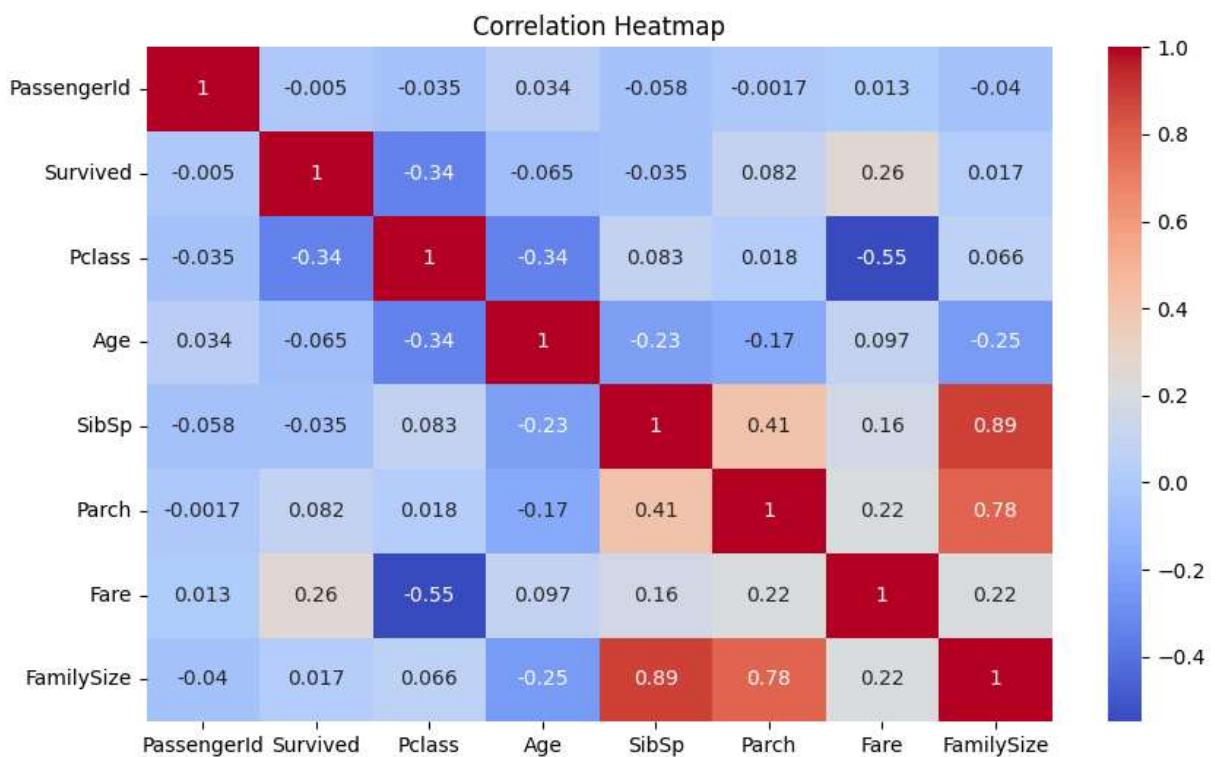


```
In [28]: sns.countplot(x='FamilySize', hue='Survived', data=df)
plt.title("Family Size VS Survival")
plt.show()
```

## Family Size VS Survival



```
In [29]: plt.figure(figsize=(10,6))
sns.heatmap(df.corr(numeric_only=True), annot=True, cmap='coolwarm')
plt.title("Correlation Heatmap")
plt.show()
```



In [30]: df.head()

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.250C
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...)	female	38.0	1	0	PC 17599	71.2833
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9250
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.100C
4	5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.050C



In [ ]: