In [1]:
```
import pandas as pd
import numpy as np
```

In [3]:
```
# Load the dataset
df = pd.read_csv('Desktop/prodigy/titanic/titanic.csv')
```

In [4]:
```
# Display the first few rows
df.head()
```

Out[4]:

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	NaN
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.2833	C85
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9250	NaN
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1000	C123
4	5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.0500	NaN

In [5]:
```
#**Data Cleaning**
#checking missing values
df.isnull().sum()
```

Out[5]:
```
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 [6]:
```
# Handle missing values (e.g., impute mean for numerical features, drop rows for ca
df.dropna(subset=["Embarked"], inplace=True)
df["Cabin"].fillna("Unknown", inplace=True)
df["Age"].fillna(df["Age"].mean(), inplace=True)
```

In [7]:
```
df.isna().sum()
```

Out[7]:
```
PassengerId    0
Survived       0
Pclass         0
Name           0
Sex            0
Age            0
SibSp          0
Parch          0
Ticket         0
Fare           0
Cabin          0
Embarked       0
dtype: int64
```

In [8]:
```
#checking duplicates
df.duplicated().sum()
```

Out[8]:
```
0
```

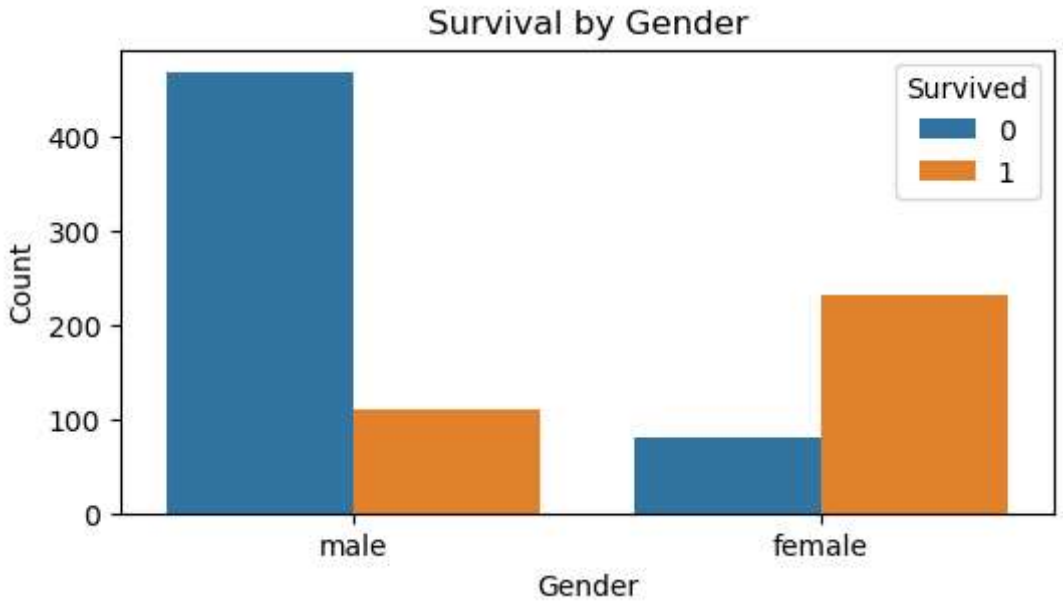
In [9]:
```
df.describe()
```

Out[9]:

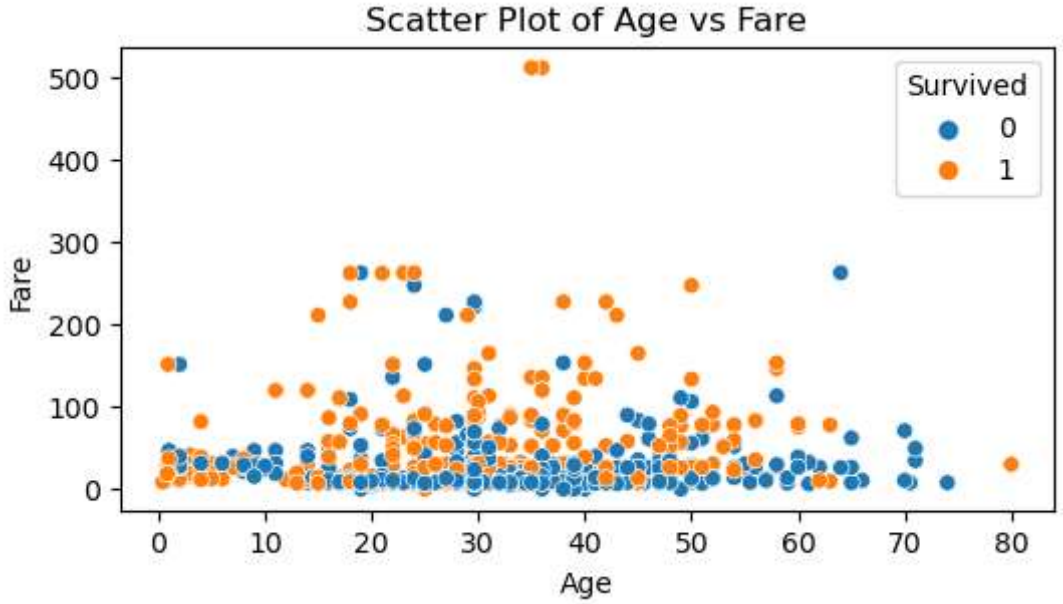
	PassengerId	Survived	Pclass	Age	SibSp	Parch	Fare
count	889.000000	889.000000	889.000000	889.000000	889.000000	889.000000	889.000000
mean	446.000000	0.382452	2.311586	29.642093	0.524184	0.382452	32.096681
std	256.998173	0.486260	0.834700	12.968346	1.103705	0.806761	49.697504
min	1.000000	0.000000	1.000000	0.420000	0.000000	0.000000	0.000000
25%	224.000000	0.000000	2.000000	22.000000	0.000000	0.000000	7.895800
50%	446.000000	0.000000	3.000000	29.642093	0.000000	0.000000	14.454200
75%	668.000000	1.000000	3.000000	35.000000	1.000000	0.000000	31.000000
max	891.000000	1.000000	3.000000	80.000000	8.000000	6.000000	512.329200

In [10]:
```
import seaborn as sns
import matplotlib.pyplot as plt
```

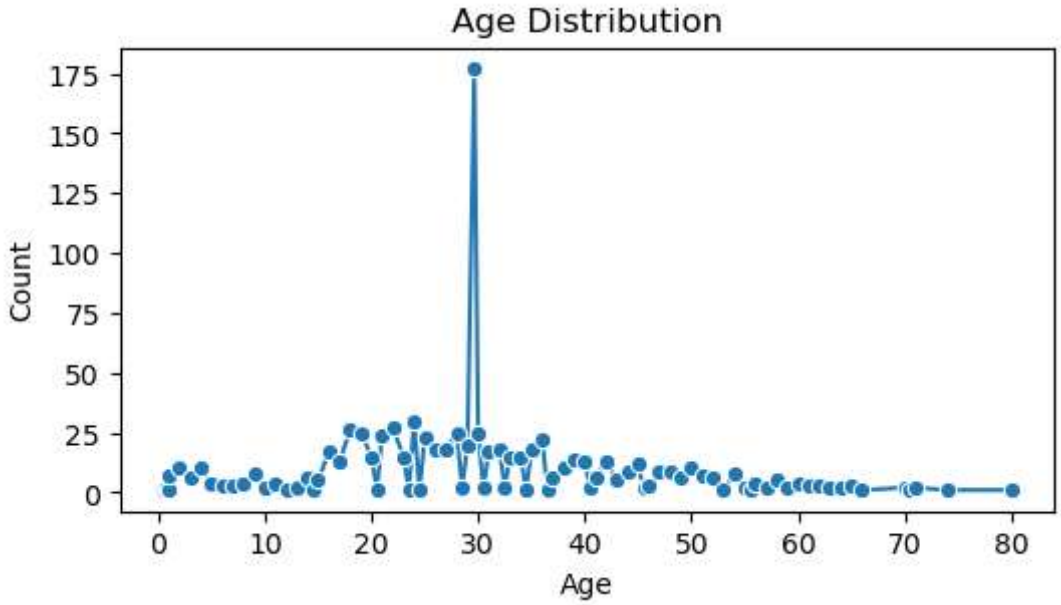
In [15]:
```
plt.figure(figsize=(6, 3))
sns.countplot(df, x="Sex", hue="Survived")
plt.title("Survival by Gender")
plt.xlabel("Gender")
plt.ylabel("Count")
plt.legend(title="Survived", loc="upper right")
plt.show()
```



In [14]:
```
plt.figure(figsize=(6, 3))
sns.scatterplot(df, x="Age", y="Fare", hue="Survived")
plt.title("Scatter Plot of Age vs Fare")
plt.xlabel("Age")
plt.ylabel("Fare")
plt.legend(title="Survived")
plt.show()
```



In [17]:
```
plt.figure(figsize=(6, 3))
sns.lineplot(data=df["Age"].value_counts().sort_index(), marker='o')
plt.title("Age Distribution")
plt.xlabel("Age")
plt.ylabel("Count")
plt.show()
```


Age Distribution
Count
Age
0
25
50
75
100
125
150
175
0
10
20
30
40
50
60
70
80