In [3]:
```
import pandas as pd

url = "https://raw.githubusercontent.com/datasciencedojo/datasets/master/titanic.cs
df = pd.read_csv(url)
```

df.head()

In [7]:
```
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()
```

Out[9]:

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

In [11]:
```
df['Survived'].value_counts()
```

Out[11]:
```
Survived
0    549
1    342
Name: count, dtype: int64
```

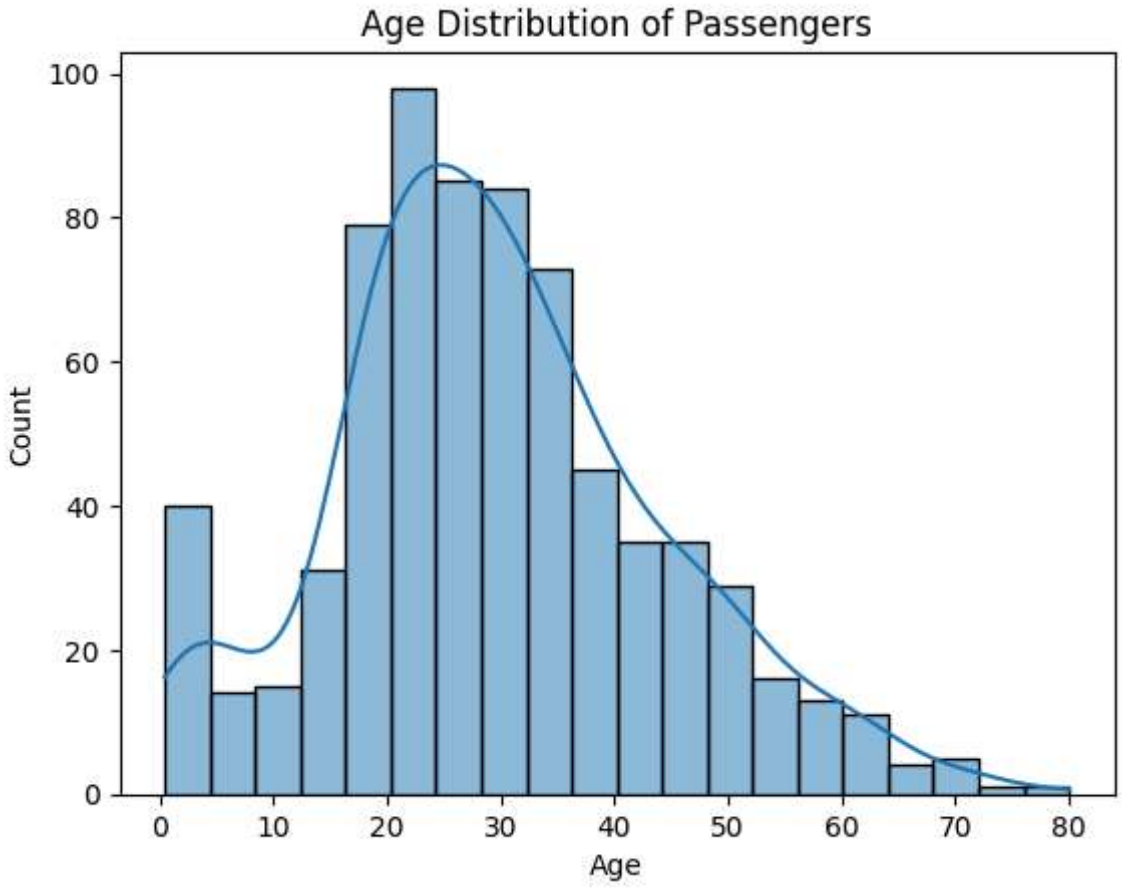
In [13]:
```
df.isnull().sum()
```

Out[13]:
```
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
```

Observations from Basic EDA

- The dataset contains missing values in the Age and Cabin columns.
- The target variable 'Survived' is imbalanced.
- The dataset includes both numerical and categorical features.

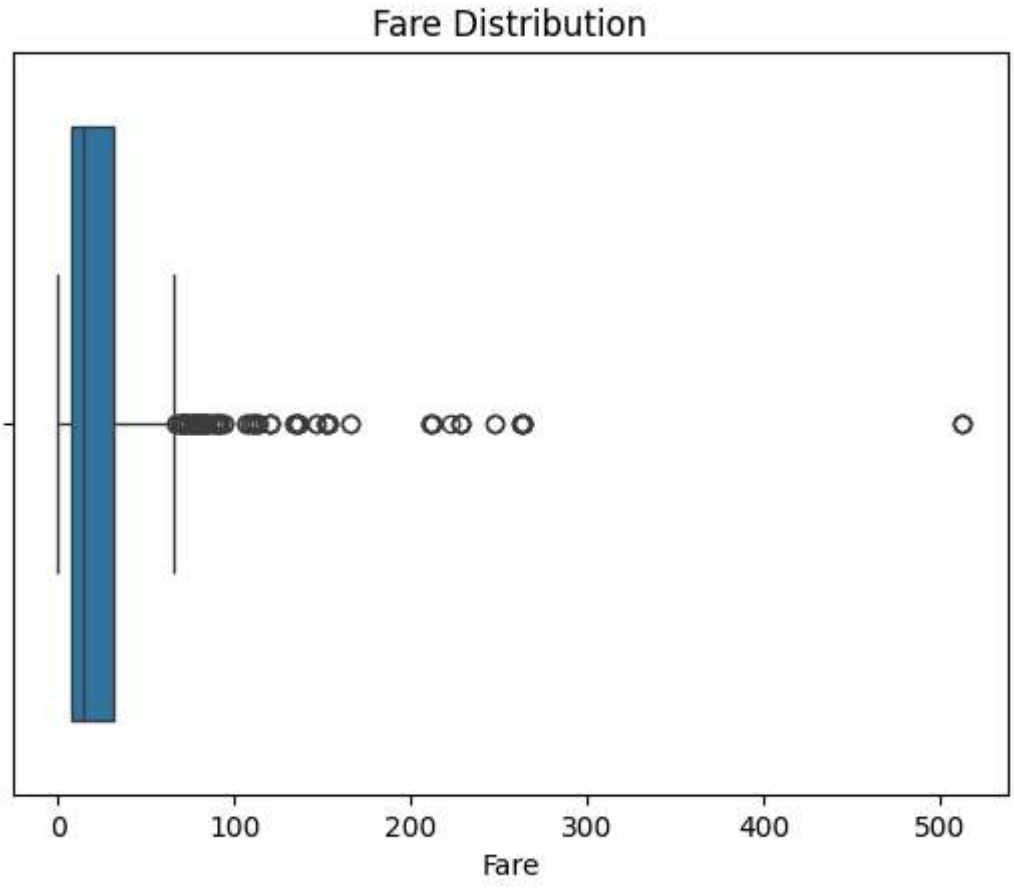
In [18]:
```
import matplotlib.pyplot as plt
import seaborn as sns

sns.histplot(df['Age'], kde=True)
plt.title("Age Distribution of Passengers")
plt.show()
```



Observation:

- The age distribution is right-skewed.
- Most passengers are between 20 and 40 years old.

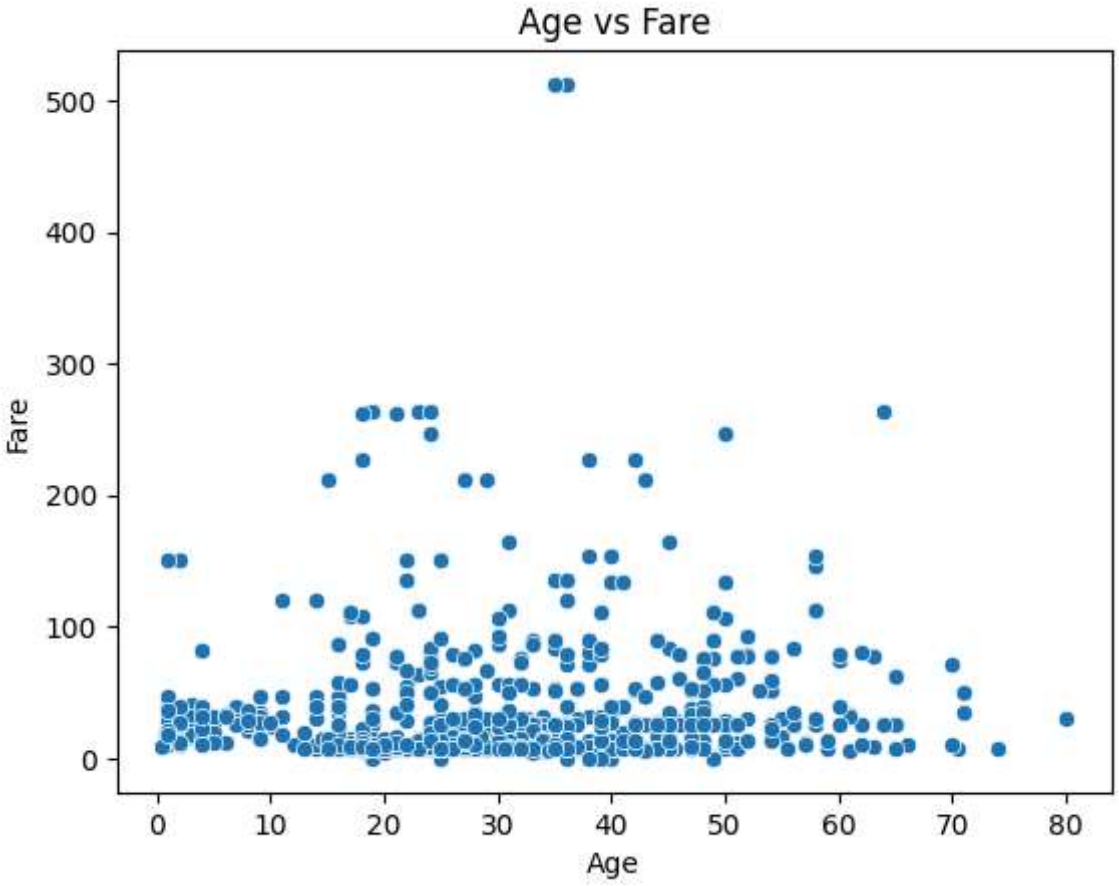
In [21]:
```
sns.boxplot(x=df['Fare'])
plt.title("Fare Distribution")
plt.show()
```



Observation:

- Fare values are highly skewed.
- Presence of outliers indicates some passengers paid very high fares.

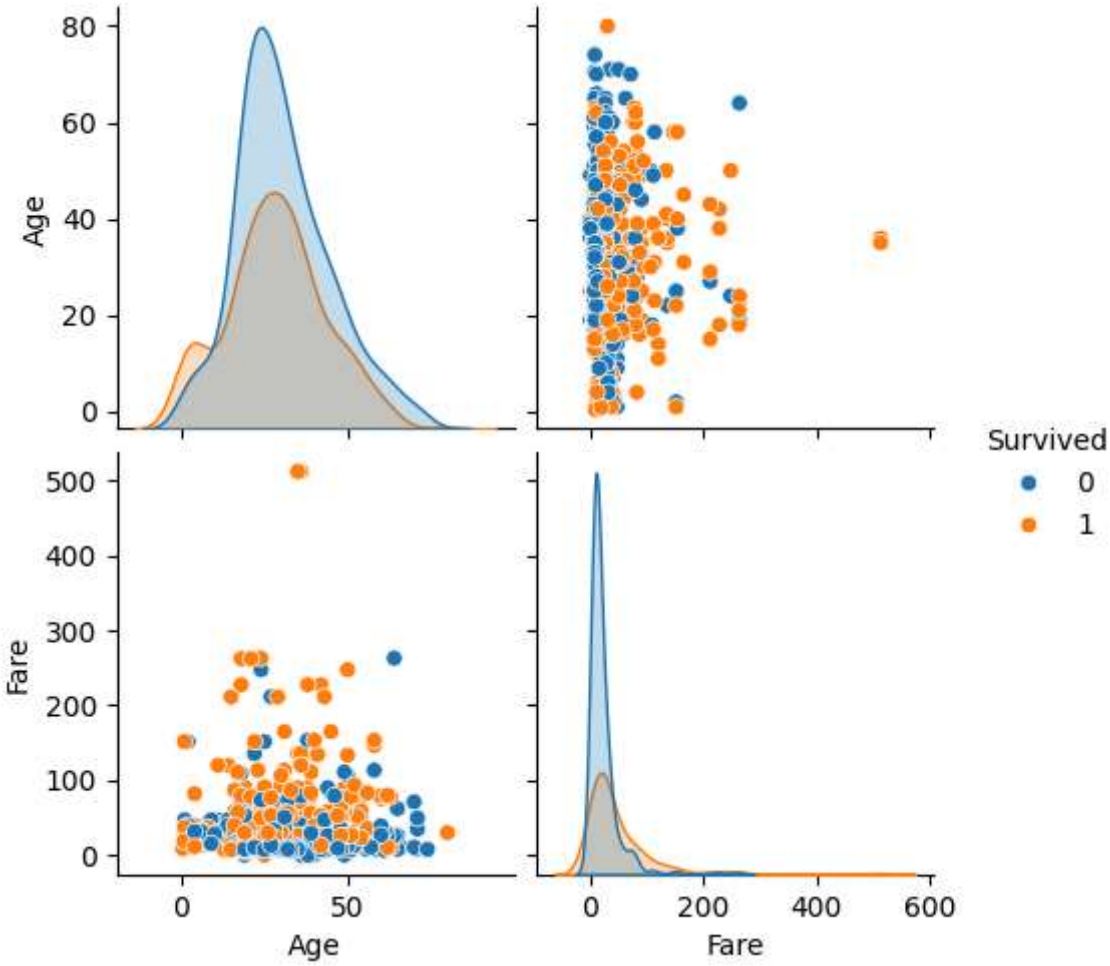
In [24]:
```
sns.scatterplot(x='Age', y='Fare', data=df)
plt.title("Age vs Fare")
plt.show()
```



Observation:

- There is no strong linear relationship between Age and Fare.
- Higher fares are spread across different age groups.

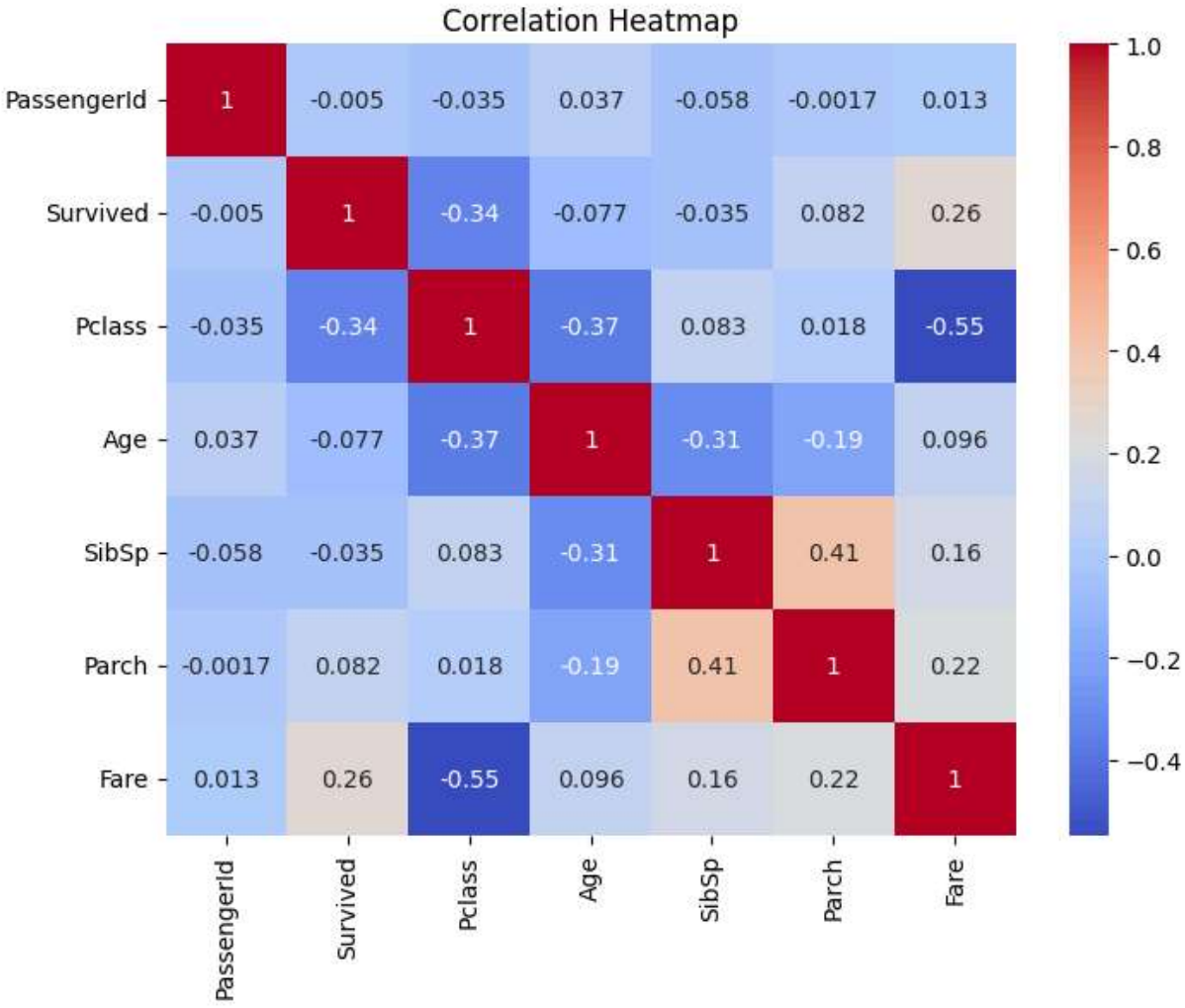
In [30]:
```
import seaborn as sns
import matplotlib.pyplot as plt

sns.pairplot(df[['Age', 'Fare', 'Survived']], hue='Survived')
plt.show()
```



Observation:

- Survival shows some relationship with Fare.
- Passengers who paid higher fares had better survival chances.
- Age alone does not strongly separate survival outcomes.

In [33]:
```
plt.figure(figsize=(8,6))
sns.heatmap(df.corr(numeric_only=True), annot=True, cmap='coolwarm')
plt.title("Correlation Heatmap")
plt.show()
```



Observation:

- Fare has a positive correlation with survival.
- Pclass shows a negative correlation with survival.
- No strong multicollinearity is observed among numerical features.

Summary of Findings

1. The Titanic dataset contains missing values in Age and Cabin columns.
2. Survival rate is influenced by passenger class and fare.
3. Higher fare passengers had better survival chances.
4. Age does not show a strong correlation with survival.
5. The dataset shows no severe multicollinearity among numeric variables.

In [ ]: