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

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

df.head()
```

Out[1]:

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

In [2]:
```
print(df.describe())
print(df.info())
print(df['Sex'].value_counts())
print(df['Embarked'].value_counts())
```

```
       PassengerId    Survived      Pclass         Age       SibSp  \
count   891.000000  891.000000  891.000000  714.000000  891.000000
mean    446.000000    0.383838    2.308642   29.699118    0.523008
std     257.353842    0.486592    0.836071   14.526497    1.102743
min       1.000000    0.000000    1.000000    0.420000    0.000000
25%     223.500000    0.000000    2.000000   20.125000    0.000000
50%     446.000000    0.000000    3.000000   28.000000    0.000000
75%     668.500000    1.000000    3.000000   38.000000    1.000000
max     891.000000    1.000000    3.000000   80.000000    8.000000

            Parch        Fare
count  891.000000  891.000000
mean     0.381594   32.204208
std      0.806057   49.693429
min      0.000000    0.000000
25%      0.000000    7.910400
50%      0.000000   14.454200
75%      0.000000   31.000000
max      6.000000  512.329200
<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
male      577
female    314
Name: Sex, dtype: int64
S    644
C    168
```

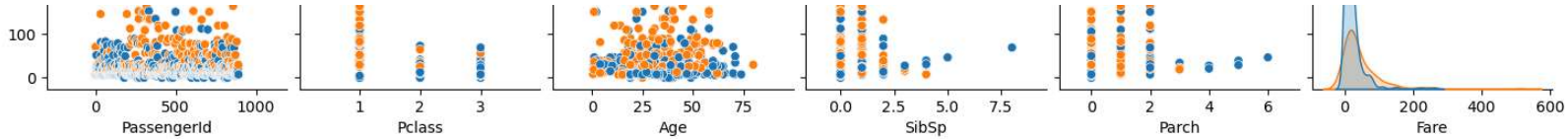
```
Q     77
Name: Embarked, dtype: int64
```

In [3]:
```
import seaborn as sns

sns.pairplot(df, hue="Survived")
```

Out[3]:
```
<seaborn.axisgrid.PairGrid at 0x175bb35e680>
```

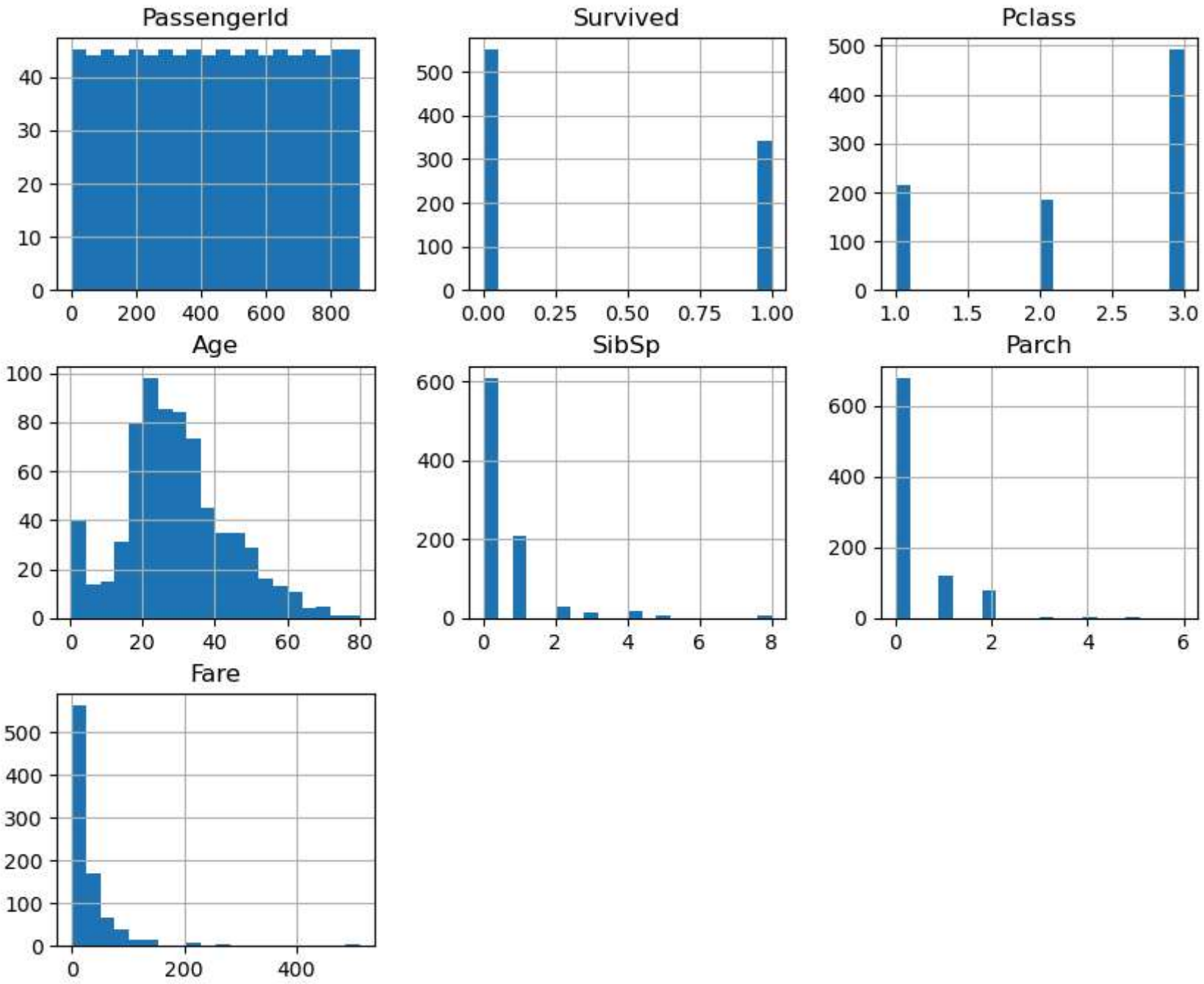

PassengerId
Pclass
Age
SibSp
Parch
Fare
Survived
0
1



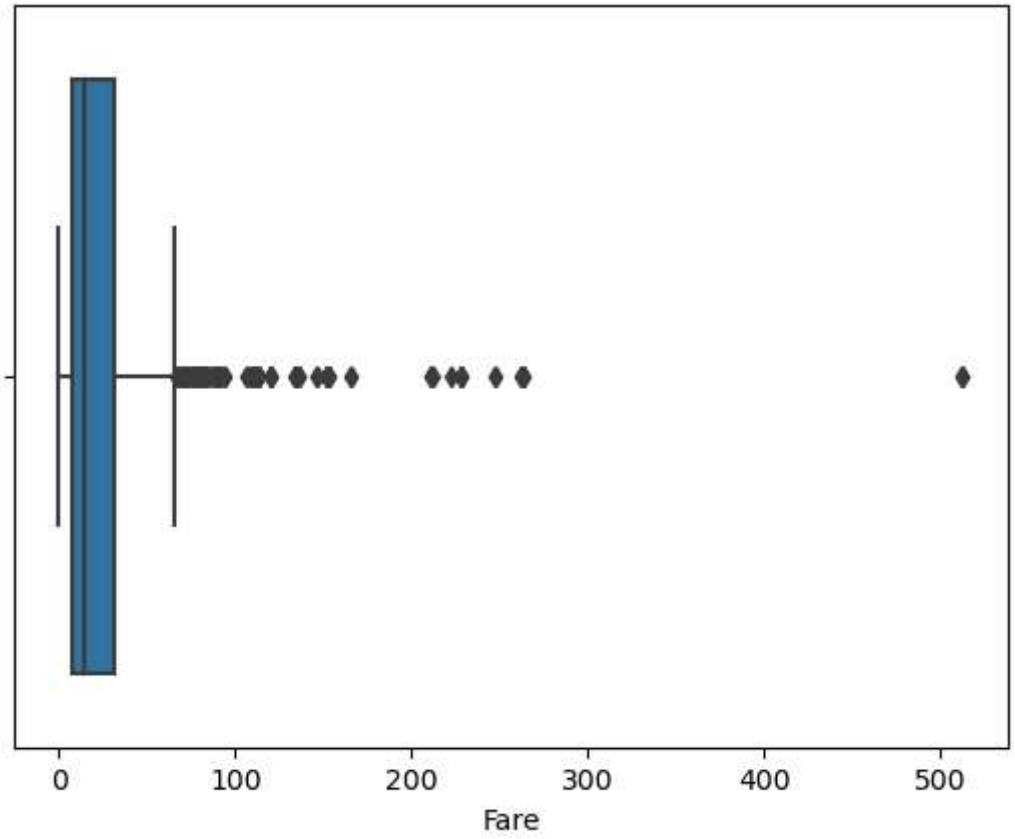
In [4]:
```
import matplotlib.pyplot as plt
corr_matrix = df.corr()
plt.figure(figsize=(10,6))
sns.heatmap(corr_matrix, annot=True, cmap="coolwarm", fmt=".2f")
plt.show()
```

```
C:\Users\kathi\AppData\Local\Temp\ipykernel_32792\2382276771.py:2: FutureWarning: The default value of numeric_only in
DataFrame.corr is deprecated. In a future version, it will default to False. Select only valid columns or specify the v
alue of numeric_only to silence this warning.
  corr_matrix = df.corr()
```

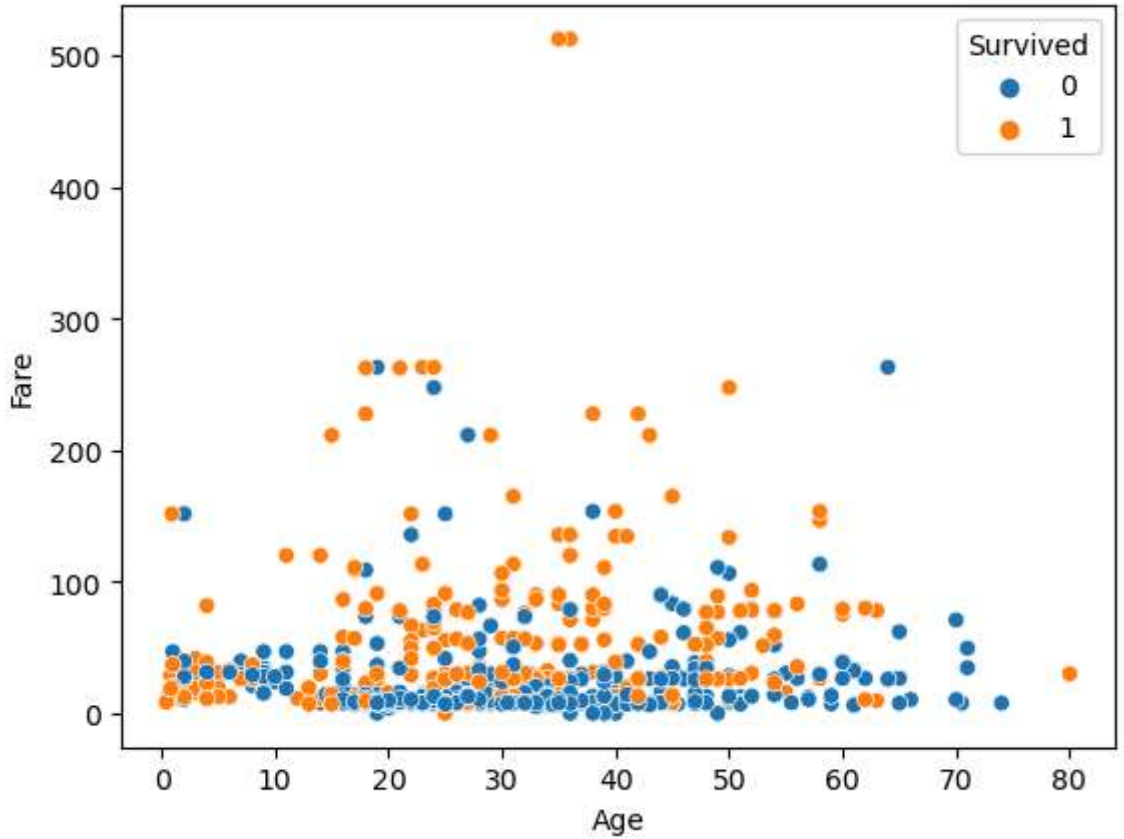
In [5]:
```
df.hist(figsize=(10,8), bins=20)
plt.show()
```



In [6]:
```
sns.boxplot(x=df["Fare"])
plt.show()
```



In [7]:
```
sns.scatterplot(x=df["Age"], y=df["Fare"], hue=df["Survived"])
plt.show()
```



Summary of Findings:

->The dataset contains demographic and travel information of Titanic passengers, including survival status.

->Female passengers had a higher survival rate compared to males. This was evident in the visual distribution and value counts.

->Most passengers were between 20 to 40 years of age, based on the histogram of the Age column.

->Passengers who paid higher fares had a slightly better chance of survival. The scatterplot between Age and Fare showed that many survivors were clustered in the lower age and higher fare group.

->The heatmap showed moderate correlation between Fare, Pclass, and Survived, suggesting these features had predictive value.

->The pairplot revealed that survival was more frequent among passengers from 1st class and younger age groups.

->The boxplot on Fare showed presence of outliers with very high ticket prices, typically associated with 1st class.