In [5]:
```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
```

In [17]:
```
train_df = pd.read_csv("E:/Elevate Labs Internship/Task 5/train.csv")
```

In [19]:
```
train_df.head()
```

Out[19]:

	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	7.925C
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 [21]:
```
gender_submission_df = pd.read_csv("E:/Elevate Labs Internship/Task 5/gender_submis
```

In [23]:
```
gender_submission_df.head()
```

Out[23]:

	PassengerId	Survived
0	892	0
1	893	1
2	894	0
3	895	0
4	896	1

In [25]:
```
test_df= pd.read_csv("E:/Elevate Labs Internship/Task 5/test.csv")
```

In [27]:
```
test_df.head()
```

Out[27]:

	PassengerId	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Em
0	892	3	Kelly, Mr. James	male	34.5	0	0	330911	7.8292	NaN	
1	893	3	Wilkes, Mrs. James (Ellen Needs)	female	47.0	1	0	363272	7.0000	NaN	
2	894	2	Myles, Mr. Thomas Francis	male	62.0	0	0	240276	9.6875	NaN	
3	895	3	Wirz, Mr. Albert	male	27.0	0	0	315154	8.6625	NaN	
4	896	3	Hirvonen, Mrs. Alexander (Helga E Lindqvist)	female	22.0	1	1	3101298	12.2875	NaN	

In [29]:
```
train_df.describe()
```

Out[29]:

	PassengerId	Survived	Pclass	Age	SibSp	Parch	Fare
count	891.000000	891.000000	891.000000	714.000000	891.000000	891.000000	891.00000(
mean	446.000000	0.383838	2.308642	29.699118	0.523008	0.381594	32.20420ٍ8
std	257.353842	0.486592	0.836071	14.526497	1.102743	0.806057	49.69342(
min	1.000000	0.000000	1.000000	0.420000	0.000000	0.000000	0.00000(
25%	223.500000	0.000000	2.000000	20.125000	0.000000	0.000000	7.91040(
50%	446.000000	0.000000	3.000000	28.000000	0.000000	0.000000	14.45420(
75%	668.500000	1.000000	3.000000	38.000000	1.000000	0.000000	31.00000(
max	891.000000	1.000000	3.000000	80.000000	8.000000	6.000000	512.32920(

In [31]:
```
train_df.shape
```

Out[31]: (891, 12)

In [33]:
```
train_df.loc[:,['Survived', 'Sex']]
```

Out[33]:

	Survived	Sex
0	0	male
1	1	female
2	1	female
3	1	female
4	0	male
...	...	...
886	0	male
887	1	female
888	0	female
889	1	male
890	0	male

891 rows × 2 columns

In [35]:
```
train_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 [37]:
```
train_df.count()
```

Out[37]:
```
PassengerId    891
Survived       891
Pclass         891
Name           891
Sex            891
Age            714
SibSp          891
Parch          891
Ticket         891
Fare           891
Cabin          204
Embarked       889
dtype: int64
```

In [39]:
```
train_df['PassengerId'].isna().sum()
```

Out[39]:
```
0
```

In [41]:
```
train_df['Embarked'].isna().sum()
```

Out[41]:
```
2
```

In [43]:
```
train_df['Age'].isna().sum()
```

Out[43]:
```
177
```

In [45]:
```
train_df.value_counts()
```

Out[45]:
```
PassengerId  Survived  Pclass  Name
Sex     Age   SibSp  Parch  Ticket    Fare      Cabin  Embarked
2            1         1       Cumings, Mrs. John Bradley (Florence Briggs Thayer)
female  38.0  1      0      PC 17599  71.2833   C85    C            1
572          1         1       Appleton, Mrs. Edward Dale (Charlotte Lamson)
female  53.0  2      0      11769     51.4792   C101   S            1
578          1         1       Silvey, Mrs. William Baird (Alice Munger)
female  39.0  1      0      13507     55.9000   E44    S            1
582          1         1       Thayer, Mrs. John Borland (Marian Longstreth Morri
s)  female  39.0  1      1      17421     110.8833  C68    C            1
584          0         1       Ross, Mr. John Hugo
male    36.0  0      0      13049     40.1250   A10    C            1

..
328          1         2       Ball, Mrs. (Ada E Hall)
female  36.0  0      0      28551     13.0000   D      S            1
330          1         1       Hippach, Miss. Jean Gertrude
female  16.0  0      1      111361    57.9792   B18    C            1
332          0         1       Partner, Mr. Austen
male    45.5  0      0      113043    28.5000   C124   S            1
333          0         1       Graham, Mr. George Edward
male    38.0  0      1      PC 17582  153.4625  C91    S            1
890          1         1       Behr, Mr. Karl Howell
male    26.0  0      0      111369    30.0000   C148   C            1
Name: count, Length: 183, dtype: int64
```

In [47]:
```
train_df['Survived'].count()
```

Out[47]: 891

In [49]:
```
train_df['Age'].aggregate(['max', 'min'])
```

Out[49]:
```
max    80.00
min     0.42
Name: Age, dtype: float64
```

In [51]:
```
train_df['Sex'].value_counts()
train_df['Embarked'].value_counts()
```

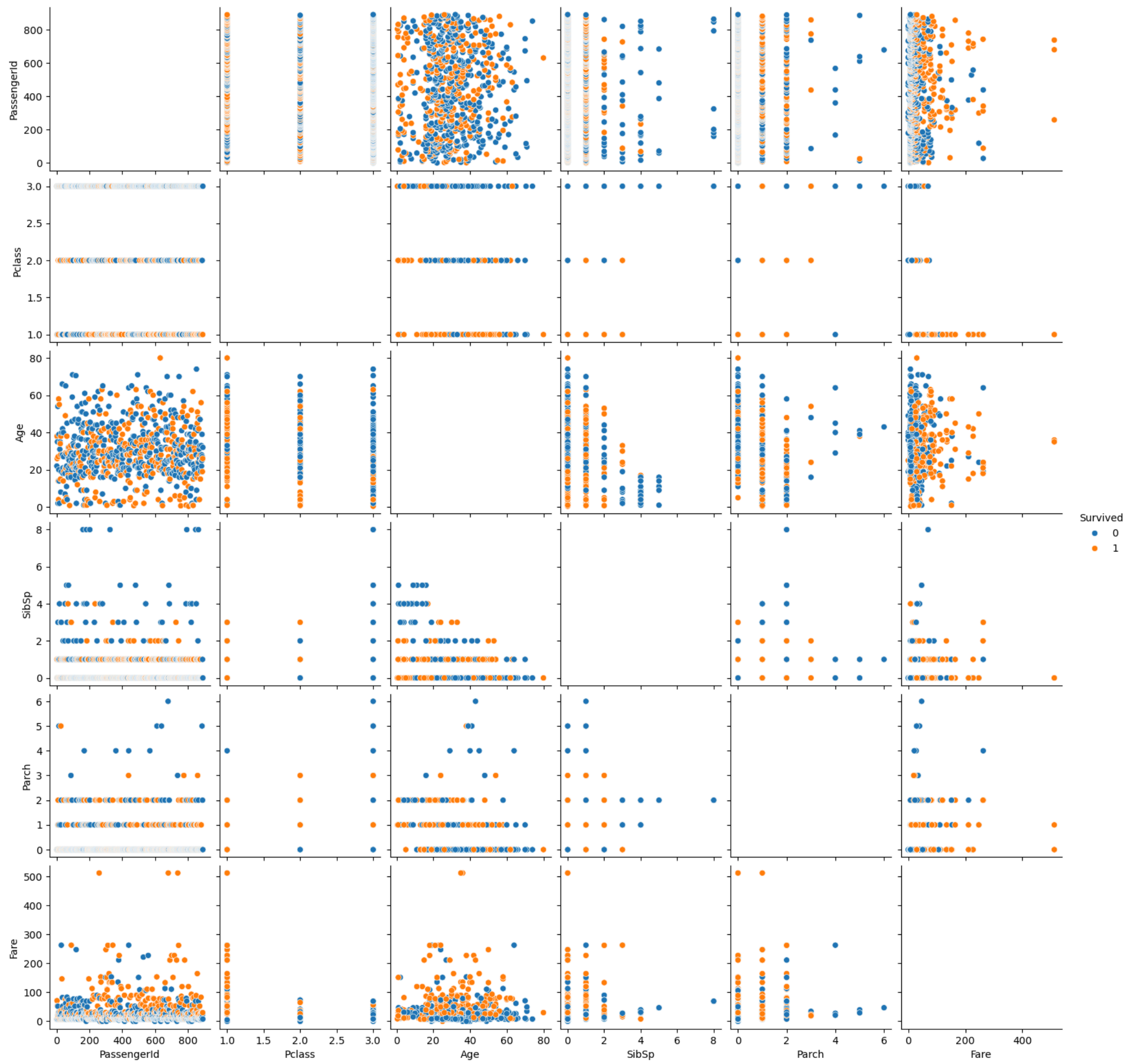
Out[51]:
```
Embarked
S    644
C    168
Q     77
Name: count, dtype: int64
```

In [69]:
```
train_df['Survived'].value_counts()
train_df['Sex'].value_counts()
```

Out[69]:
```
Sex
male      577
female    314
Name: count, dtype: int64
```

In [55]:
```
sns.pairplot(train_df, hue="Survived", diag_kind="histogram")
```

Out[55]:
```
<seaborn.axisgrid.PairGrid at 0x258e8a7e4e0>
```



Pclass vs. Age: Passengers in 1st class (lower number = higher class) tend to be older. More survivors (orange) are seen in 1st class, while non-survivors are concentrated in 3rd class. Indicates a strong relationship between Pclass and survival.

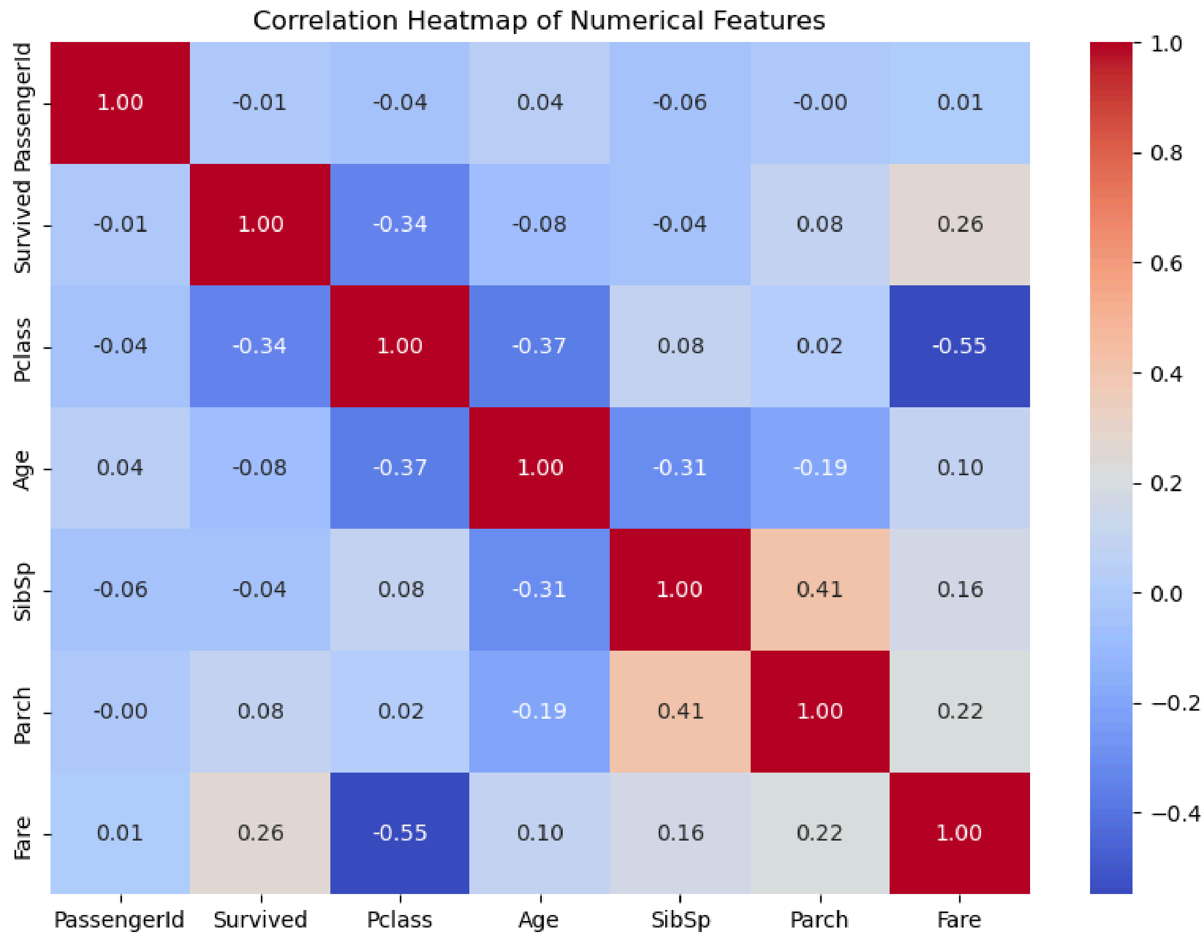
Fare vs. Pclass: A clear trend shows higher fares for higher classes, especially 1st class. Survivors are more concentrated in the higher fare bracket, suggesting people who paid more had better survival chances.

Age vs. Survival: There's a wide age range for both survivors and non-survivors. Children (under ~10 years) had a better survival rate. Adults have mixed survival chances; no very strong direct pattern here without further breakdown (like gender or class).

SibSp (siblings/spouses aboard) and Parch (parents/children aboard): Most passengers had 0 or 1 family member aboard. People traveling with small families (1–2) may have had a slightly better survival rate. Large family sizes (higher SibSp/Parch) show less survival—could be due to difficulty escaping together.

Fare vs. Age: Older passengers generally paid higher fares, especially in 1st class. Survival is higher in older passengers who paid more, again suggesting socio-economic advantage.

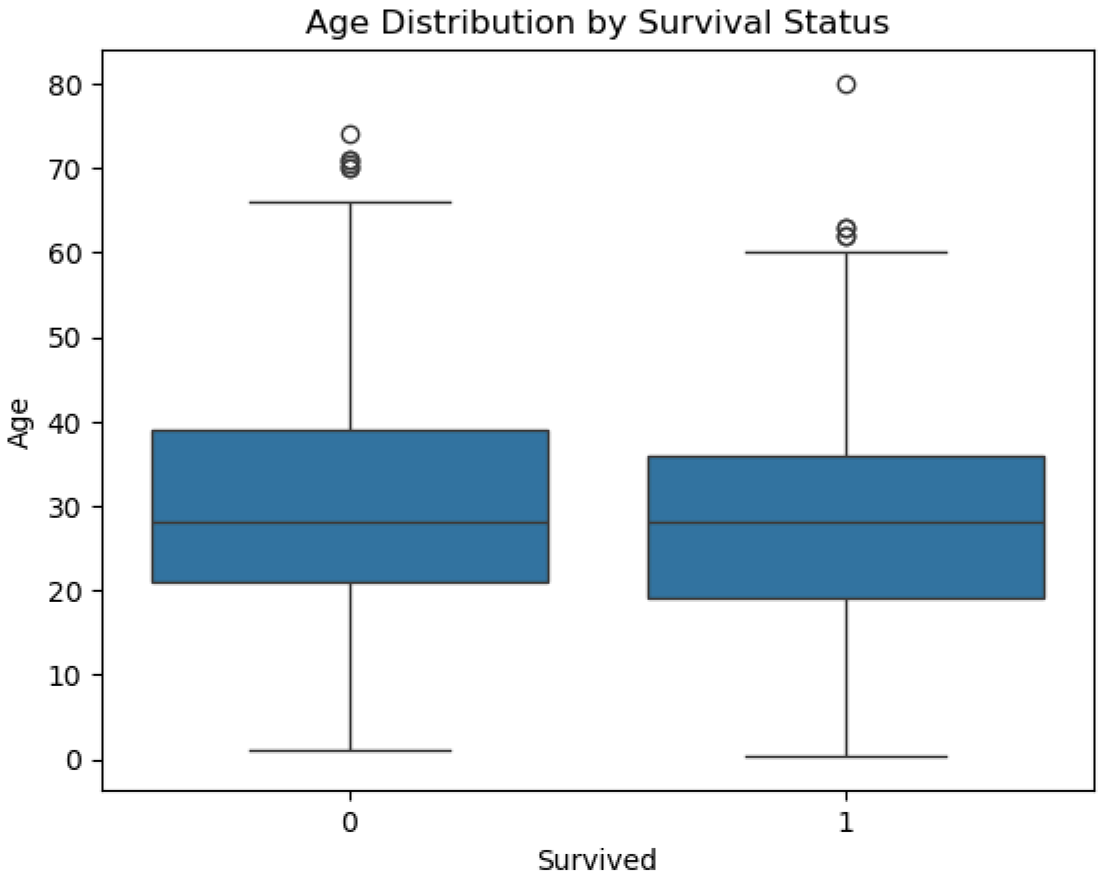
In [59]:
```
# Select only numeric columns
numeric_df = train_df.select_dtypes(include=['number'])

# Now compute correlation
corr = numeric_df.corr()
plt.figure(figsize=(10, 7))
sns.heatmap(corr, annot=True, cmap="coolwarm", fmt=".2f")
plt.title("Correlation Heatmap of Numerical Features")
plt.show()
```



Fare and Pclass are most strongly correlated with survival. Passengers who paid more (likely 1st class) had better chances of surviving. **Pclass** has a **moderate negative correlation** with survival (`-0.34`), confirming that 3rd class passengers had lower survival rates. **Age, SibSp, and Parch** show **very weak correlation** with survival — they may still be useful when combined with other features, but not individually strong. **Fare and Pclass** are strongly negatively correlated (`-0.55`), supporting the trend of socio-economic influence on survival.

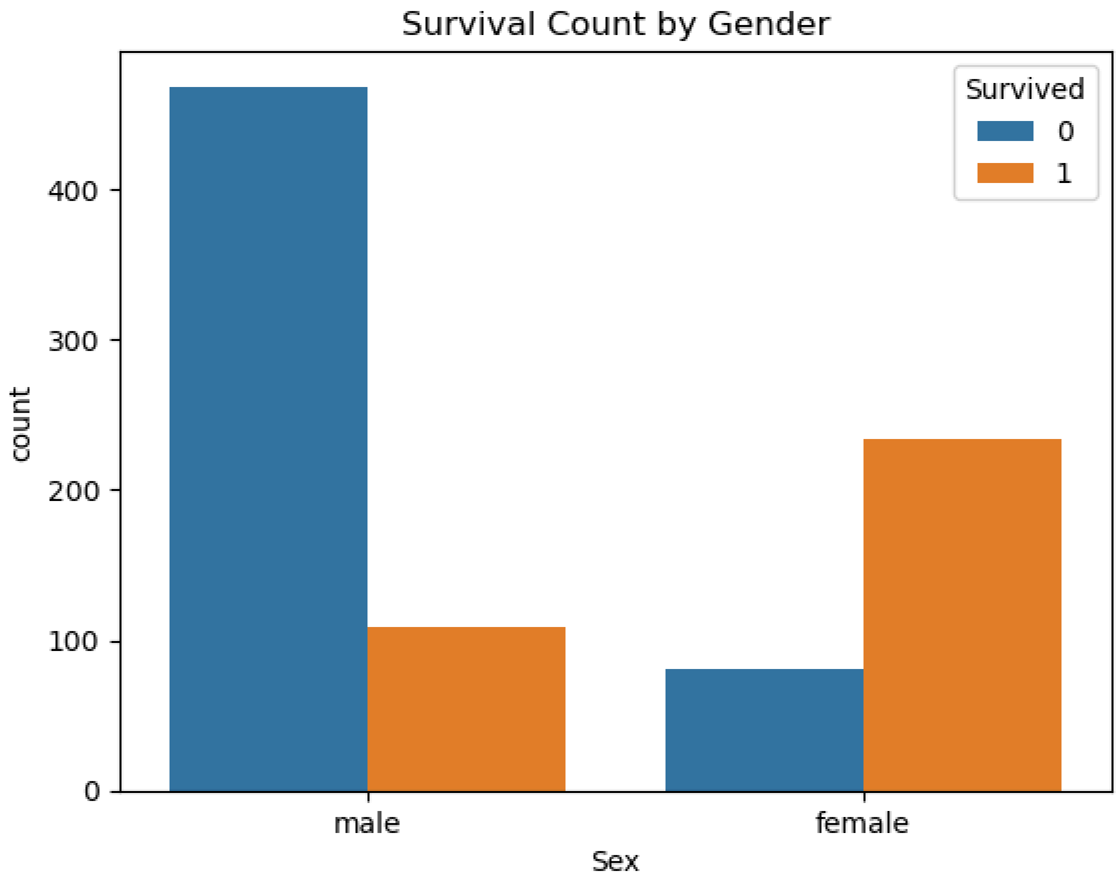
In [61]:
```
sns.boxplot(x="Survived", y="Age", data=train_df)
plt.title("Age Distribution by Survival Status")
```

```
plt.show()
```



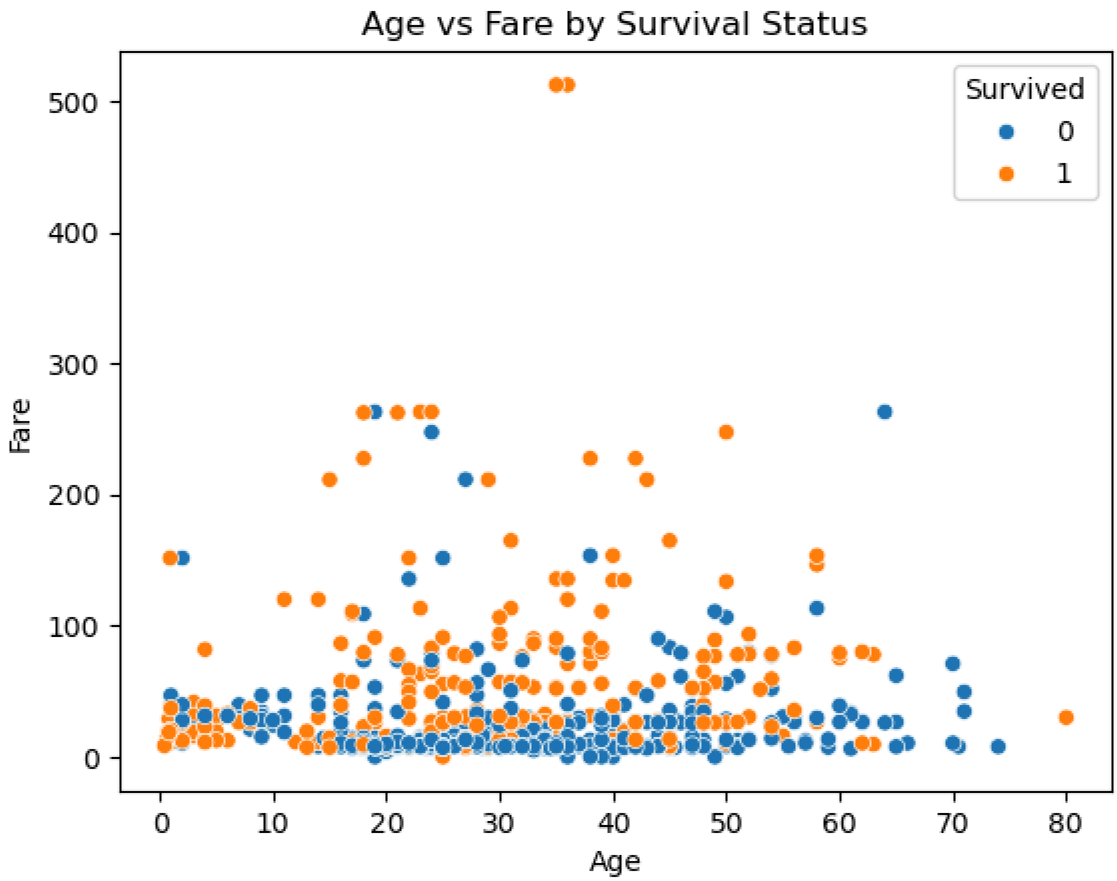
The **median age** is almost the same for survivors and non-survivors, suggesting that **age was not a strong influential factor** in survival. The **spread of ages** is slightly wider for non-survivors. A **few children** appear in the survivor group, which may reflect **priority evacuation** for younger passengers. Both groups show **outliers** among older individuals, but **more older passengers died** than survived.

In [63]:
```
sns.countplot(x='Sex', hue='Survived', data=train_df)
plt.title("Survival Count by Gender")
plt.show()
```



The plot clearly shows that **females had a significantly higher survival rate** than males. A majority of **males did not survive**, whereas **most females survived**. This supports the well-known rescue principle of "**women and children first**" applied during the Titanic disaster.

In [65]:
```
sns.scatterplot(x='Age', y='Fare', hue='Survived', data=train_df)
plt.title("Age vs Fare by Survival Status")
plt.show()
```



Passengers who **paid higher fares were more likely to survive**. Most **non-survivors paid low fares**, likely indicating they were in lower classes. **Age does not show a strong pattern** of correlation with survival in this plot. This suggests that **economic status (fare)** might have been a more important factor than age for survival.

Summary of EDA Findings (Titanic Dataset)

Correlation Heatmap: Fare shows a positive correlation (0.26) with Survival, indicating passengers who paid higher fares were more likely to survive. Pclass has a negative correlation (-0.34) with Survival, meaning passengers in higher classes (1st class) had higher survival chances. Other numerical features like Age, SibSp, and Parch have relatively weak correlations with survival.

Age Distribution by Survival (Boxplot): The median age of survivors and non-survivors is similar. Both groups have a wide age range, but survivors show slightly lower variability. No strong pattern suggesting that age alone determined survival.

Survival Count by Gender (Countplot): Female passengers had a significantly higher survival rate. Most male passengers did not survive, while most female passengers survived. This supports the "women and children first" policy during evacuation.

Age vs Fare by Survival (Scatter Plot): Survivors tend to have paid higher fares, especially those aged 20–50. Non-survivors are mostly clustered around lower fares, regardless of age. Indicates that economic status (as reflected by Fare) had a stronger influence on survival than age.

In [ ]:

```

```